FACULTEIT ECONOMIE EN
BEDRIJFSWETENSCHAPPEN

KU LEUVEN

Contributions to the analysis of credit risk data using advanced survival analysis techniques.

Proefschrift Voorgedragen tot
het Behalen van de Graad van
Doctor in de Toegepaste
Economische Wetenschappen

door

Lore DIRICK

NUMMER 500 2015

Committee

Advisor:	
Prof. Dr. Gerda Claeskens	*KU Leuven*
Co-advisor:	
Prof. Dr. Bart Baesens	*KU Leuven*
Chair:	
Prof. Dr. Piet Sercu	*KU Leuven*
Members:	
Prof. Dr. Tony Bellotti	*Imperial College, London*
Prof. Dr. Ricardo Cao	*Universidade da Coruña*
Prof. Dr. Irène Gijbels	*KU Leuven*

Acknowledgments

This thesis concludes almost four years of intensive research, and is the result of the effort and both indirect and direct support of many people. I am very grateful for the financial support, provided by the Research Foundation - Flanders.

First of all, I would like to express profound gratitude to my advisor Prof. Gerda Claeskens and co-advisor Prof. Bart Baesens. Thank you for giving me the opportunity to start this PhD project four years ago. Though only having met briefly before I started, you have both shown your full support and faith in my abilities from the very beginning. Thank you both for passing on your great passion for research and guiding me through my PhD process. Gerda, thank you for always being there when I needed help, even though supervising five other (post)-doctoral students, having a busy teaching schedule, your own research to do and a family! I am still puzzled how you manage to combine all this. Bart, thank you for passing on your credit risk insights, the interesting contacts with banks and your help regarding my post-PhD job search. And yes, thank you for all the great pizza lunches!

Special thanks also go to the other members of the supervisory committee: Prof. Tony Bellotti, Prof. Ricardo Cao and Prof. Irène Gijbels. The constructive comments during the doctoral seminars thanks to their deep insights and expertise in the field led to a significantly improved quality of the thesis.

I additionally want to thank Professor Bellotti for the very interesting research stay at Imperial College in the final year of my PhD programme.

These two months in London were amazing, both on a personal and on an academic level. I really enjoyed our weekly meetings discussing the research, and it has been a great honor to collaborate for my final PhD project. Thanks to everyone in my London office for giving me such a warm welcome, especially to Zhana, who made me feel at home from the very beginning. I want to express my gratitude to FWO (Fonds voor Wetenschappelijk Onderzoek) for supporting me with a grant to make this research stay financially feasible.

My sincere appreciation goes out to Prof. Andrey Vasnev for raising the idea of introducing unobserved heterogeneity in mixture cure models, and collaborating on this project. Your remarks and help have substantially improved the second chapter of this thesis.

Furthermore, I am thankful that KU Leuven provided me with the facilities to carry out this doctoral research. The weekly organized seminars of the LSTAT group have broadened my knowledge and gave me the opportunity to regularly meet our nice colleagues from the statistics department in Heverlee. Running simulation studies and analyzing big data sets, the computer power provided by the VSC (Vlaams Supercomputer Centrum) was indispensable. Thank you for the financial and administrative support when attending several conferences in Belgium and abroad. Because of this support, I was able to finetune my own presentation skills, attended many interesting presentations and met amazing people.

For my research, real data sets from banks were essential. I am extremely grateful for all the banks who were willing to provide this data, despite the legal difficulties regarding confidentiality. The third chapter in this thesis would not have been there without your faith in me and my research.

To my colleagues of ORSTAT, it has been such a pleasure meeting you all. Jonathan and Andreas, thank you for the numerous healthy lunches at "soup away" at the beginning of my PhD. Some colleagues who were there at the start of my PhD left by now, but I want to thank them for memorable moments together, among others, my first conference (the BSS meeting in Liege): thank you Vishva, Chang, Marjolein, Steffi and Jean-Marc.

Special thanks go out to Charles, with whom I spent many coffee breaks discussing PhD and life in general. Thank you for the continuing friendship even though seperated by the North Sea. Thanks to my current colleagues Eugen, Ali, Thomas, Deniz, Peter, Ruben, Nicolas, Viktoria, Ines and Roel for being so enthusiastic organizing many social events among colleagues: watching world cup games, playing volleyball, throwing barbecues,... I will miss you and I cherish the memories of these great times together. Special thanks go out to Roel for the great chats and laughs while sharing an office. I will definitely miss you making fun of me every time I was amazed once more by the nice view from our office.

To all the Operational Research-colleagues of ORSTAT, the colleagues at the Operations Management research group and the colleagues at LIRIS: thank you for the occasional chats during coffee break and lunch. Special thanks go to Aimée, having a related PhD topic funded by the same project it was always nice to catch up about our research.

Having ambitious friends is a blessing, though not always easy when ambition moves your friends all over the country. For my long-time group of six friends, however, this has not proven to be a problem. I am so proud of what everyone has achieved and always happy when I can brag about every single one of you! Nina and Eva, sharing every step of our lives since kindergarten, I couldn't have wished for better friends to literally share my entire life with. Nina, being the person I shared my oldest memory with, I think it is amazing that we have both chosen to do a PhD after our studies. It was great to talk about academics and parallels on our workfloor, even though the differences in our fields. Eva, thank you for sharing our love for music. From making "musicals" together when we were in primary school to playing music together on weddings in recent years, these are the moments I cherish. Ilse, Marie and Floor, thank you for joining our group of three in high school. I highly value the memories of all our trips together, and hope many more are to come in the future. Ilse, thank you for being the only one keeping me company in Leuven. The Scrabble and beer nights were welcome distractions after an exhausting day of research.

I want to thank Linde, along with tons of knowledge and international

experience the most valuable product of my student years. I highly value the intense friendship we have built over the last few years, and am extremely grateful you were back in Belgium in the finishing months of my PhD.

To all the friends I met during my business engineering studies, the people I shared a house with, the friends of Hannes who have become my friends over the years, my quizteam, the music bands I was part of and everyone who cannot be placed in the previous categories but with whom I shared memorable moments: thank you!

Thank you Karine, Bernard, Michiel, Laurens, Roxanne and Lieze for being such an amazing family to Hannes, and by extension, myself. Thank you for many nice dinners, helping us move all the time (yes, it happend quite a lot!) and the nice weekends in Oostende and London.

Last but definitely not least I want to thank my family and boyfriend Hannes. Thank you Hannes, for always being there. Thank you for convincing me to start a PhD programme. Thank you for supporting me in every decision I took during my PhD, and in the international adventure that lays ahead. I know this is not the easiest thing to do, but I am so greatful that you are by my side. Mama and papa, I just don't know where to start. I am where I am now thanks to your unconditional love and support. You are not just my parents, but also the people I can talk to about anything, and I cannot thank you enough for this. Anneleen, you are not just my sister but also my very best friend. You living in Hong Kong is not always easy, but knowing that you are happy there with Max is a big comfort to me and the rest of the family. Luckily, technology lightens the burden (thank you, Skype and Whatsapp)! I am beyond happy that you are coming to Belgium especially for my defense. Dries, thank you for being the funniest and smartest brother. I am so happy you are having a blast on your exchange semester in Australia while I am writing this thesis, and am proud your graduation is approaching as well!

Lore Dirick Leuven, September 2015.

Introduction

Since the late eighties, a set of recommendations for regulating the banking industry was being published by the Basel Committee on Banking Supervision (BCBS). This committee of banking supervisory authorities initially contained the G-10 (a group of ten countries that agreed to partake in the "General Arrangements to Borrow (GAB)") central bank governors. Although not having an explicit authority, the policies described in these Basel Accords tend to be implemented by most BCBS member countries. Currently, (among others) all G-20 major economies (an international forum to promote and pertain international financial stability) are represented in the BCBS.

The Basel Accords have changed the strategies of financial institutions significantly. The 2004 Basel II Accord, superseding the 1988 Basel I Accord, played a particularly important role in this strategy shift. One of the key goals of this accord was the encouragement of continuous improvements in risk measurement. To achieve this, the Basel II Accord stipulated that large banks should be allowed to use risk assessment based on their own models to determine the minimum amount of capital they need to hold as a buffer against unexpected losses. The credit crisis in 2008 led to substantial criticism on the Basel II Accord and this statement in particular, as many banks were being said to have entered the crisis with insufficient liquidity buffers. The need for more strict regulations became apparent, and the requisite of minimum capital requirements was supplemented by other forms of reserves in the Basel III Accord, in this way complementing the Basel II accord rather than superseding it.

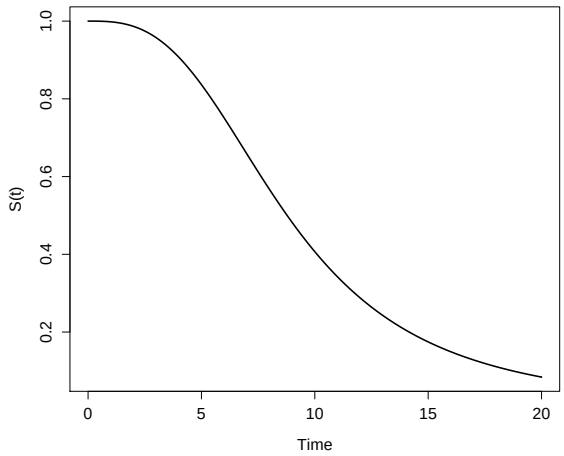

Figure 1: *A survival function*

Notwithstanding the criticism, the importance of Basel II should not be underestimated, as this accord gave rise to a more model-based focus in the banking industry. With typically a major focus on basic regression techniques such as logistic regression for modelling "good" versus "bad" customers, growing research in other areas of statistics and machine learning, which were known but less explored in this context, was a result. One of these areas, and at the same time the backbone of this dissertation, is survival analysis.

With important initial applications in actuarial sciences through life tables, survival analysis is a widespread method in the biomedical context and deals with the analysis of the duration time until a certain event, such as the time of death in biological organisms. Other fields where survival analysis is used are engineering (where it is often referred to as "reliability analysis") and sociology ("event history analysis"). A typical property of survival analysis is that this method can deal with censored data. Recapturing the biological example, censoring takes place when death is not observed for certain subjects in the sample. Typically, a survival function or curve is drawn (See Figure 1). This function represents $S(t) = P(T > t)$,

or in words, the probability that the time of the event of interest is later then some specified time t, for every t.

Defining loan "default" (or, as an extension, "early repayment") as the event of interest, the appropriateness of using survival analysis in the credit risk content becomes apparent. The advantage of using this method in this context, as opposed to more frequently used classification techniques, is that

(1) It is possible to compute a "probability of default" or a PD-estimate, which is a key parameter in credit scoring, at every point in time during the loan term.

(2) One can predict the expected time of default (more information on this can be found in Chapter 3).

Despite the fact that there are certain analogies between medical survival and survival in a credit loan context, there are also differences that might make the standard survival approach inappropriate for the analysis of credit data. The main problem is that typically, a very high proportion of credit data is right-censored, not only because the customer default is not observed during the observation period, but simply because default does not take place in the entire loan lifetime. To model a so-called "insusceptible" part of the loan population, mixture cure models can be used. A survival curve using a mixture cure model with cured fraction 0.3 is shown in Figure 2. Where a non-mixture survival function goes to zero when $t \to \infty$, the survival function goes to the cured fraction when using a mixture cure model. Although survival models in a non-mixture context make their appearance in this thesis (in Chapter 3), mixture cure models play a central role in every chapter.

In Chapter 1, the mixture cure model, for single and multiple events, is explained in more detail. As there is typically a missing data-problem, as there is no complete information on which part of the population is "susceptible" to default (or early repayment) and which part is not, an appropriate version of Akaike's information criterion (AIC) is derived and applied to these models.

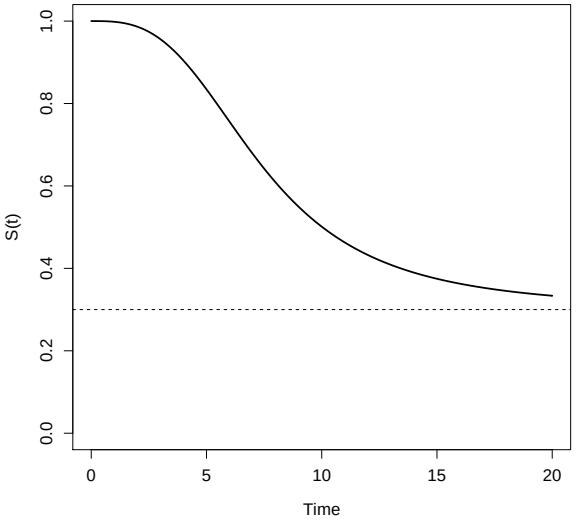

Figure 2: *A survival function using a mixture cure model*

Certain loan applicant characteristics that would affect the time of default or early repayment, might not be observed. In Chapter 2, this problem is addressed by incorporating "unobserved heterogeneity" in the mixture cure model. For model fitting purposes, a hierarchical expectation-maximization algorithm is derived.

In Chapter 3, we take a step outside the mixture cure framework, and perform a benchmark study comparing several survival techniques (both mixture and non-mixture survival models). These survival analysis techniques are applied to ten different data sets from five financial institutions, and evaluated using three different types of evaluation metrics: receiver operating characteristics curves, default time prediction and expected future values of the loan.

Standard mixture cure models include time-independent covariates in the survival analysis part of the model. Chapter 4 extends the mixture cure model such that time-dependent covariates can be included. The method is applied to real life credit data including both personal (time-independent)

information of the loan applicants and macro-economic factors that change over time.

The various chapters in this thesis can be found in:

(i) Dirick, L., Claeskens, G. and Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, **241**:449–457.

(ii) Dirick, L., Claeskens, G., Vasvnev, A. and Baesens, B. (2015). A hierarchical mixture cure model with unobserved heterogeneity using the EM-algorithm. Working paper, submitted.

(iii) Dirick, L., Claeskens, G. and Baesens, B. (2015). Time to default in credit scoring using survival analysis: a benchmark study. Working paper, submitted.

(iv) Dirick, L., Bellotti, T., Claeskens, G. and Baesens, B. (2015). Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. Working paper, submitted.

List of abbreviations

abbreviation	meaning
AFT	Accelerated failure time
AIC	Akaike information criterion
AICcd	Complete-data Akaike information criterion
AUC	Area under the curve
B-spline	Basis spline
BCBS	Basel committee on banking supervision
Cox PH	Cox proportional hazards
EFT	Expected future value
EM (algorithm)	Expectation-maximization (algorithm)
FV	Future value
GDP	Gross domestic product
MAD	Mean absolute deviation
MV	Macro-economic variable
P-spline	Penalized spline
PD	Probability of default
PE	Probability of early repayment
PM	Probability of maturity
ROC	Receiver operating characteristics
TVC	Time-varying covariate

Table of contents

Chapter 1

An Akaike information criterion for multiple event mixture cure models

Abstract

We derive the proper form of the Akaike information criterion for variable selection for mixture cure models, which are often fit via the expectation-maximization algorithm. Separate covariate sets may be used in the mixture components. The selection criteria are applicable to survival models for right-censored data with multiple competing risks and allow for the presence of a non-susceptible group. The method is illustrated on credit loan data, with pre-payment and default as events and maturity as the non-susceptible case and is used in a simulation study.

This chapter is based on Dirick, L., Claeskens, G. and Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, **241**:449–457.

1.1 Introduction

The topic of credit risk modeling has now become more important than ever before. The introduction of compliance guidelines such as Basel II and Basel III has a huge impact on the strategies of financial institutions nowadays. The Basel Accords aim at quantifying the minimum amount of buffer capital so as to provide a safety cushion against unexpected losses (Van Gestel and Baesens, 2008). A key credit risk parameter is the probability of default (PD) measuring the likelihood of an obligor to run into arrears on his/her credit obligation.

PD models are typically constructed using classification techniques such as logistic regression (Baesens et al., 2003). However, the timing when customers default is perhaps of even more interest to analyze since it can provide the bank with the ability to compute the profitability over a customer's lifetime and perform profit scoring. The problem statement of analyzing when customers default is commonly referred to as survival analysis (see, e.g., Bellotti and Crook, 2009). It is the purpose of this chapter to provide a valid model selection criterion for variable selection inside such survival models, specifically applied to credit risk modeling, with as particular characteristics allowing for defaults, maturity and early repayments in a mixture cure rate model and allowing for right-censored data.

In this chapter we deal with right-censored failure times in a mixture model context. This implies that there are two sources of incompleteness: (i) the right-censoring causes some of the event times to remain unobserved, it is only known that the event of interest did not yet take place, and (ii) not for all observations it is known to which component of the mixture model they belong; in fact, only when an observation is uncensored, we have this information. For this type of cure rate models no information criteria have yet been derived.

For incomplete and partially observed data, Cavanaugh and Shumway (1998) derive a version of the Akaike information criterion (AIC Akaike, 1973) that makes use of the expected complete data log-likelihood, rather

than the observed log-likelihood. They coined the name AICcd to this criterion. The use of the likelihood for the observed cases is discouraged since a comparison of this 'model' likelihood to a 'true' likelihood for the observed cases only is rarely of interest. By working with the complete data log-likelihood, and considering the Kullback-Leibler (KL) distance between the model and true data generating process for the complete data, the AICcd is able to select models, taking unobserved and latent variables into account. The method uses directly the output of the expectation-maximization (EM) algorithm (for more information on the EM-algorithm, we refer to McLachlan and Krishnan, 2007). We explain its definition and use below. For a comprehensive explanation of the AIC, see Claeskens and Hjort (2008, Chap. 2).

Similar variations on the AIC are studied by Claeskens and Consentino (2008), who use the output of an EM algorithm to define variable selection methods for models with missing covariate data in a linear regression setting and by Ibrahim et al. (2008) for missing data variable selection in generalized linear models.

For the case of right-censored data (not in a mixture), Liang and Zou (2008) work with an accelerated life time model and propose for that model a finite sample correction to the standard AIC, motivated from an exponential model with constant censoring. For parametric survival models Suzukawa et al. (2001) derive a version of the AIC taking the censoring into account, though require a non-standard estimation method for practical use. Fan and Li (2002) used a smoothly clipped absolute deviation penalty for the semiparametric Cox proportional hazard models, Hjort and Claeskens (2006) derived a focussed information criterion, while Xu et al. (2009) define an AIC based on the profile likelihood for proportional hazard mixed models, see also Donohue et al. (2011) for a related model selection approach. None of these papers made use of the EM algorithm to define the variable selection criterion, and neither did they consider mixture models.

In Section 1.2 we first consider the Akaike information criterion for the case of a mixture cure model with one event of interest and a group

non-susceptible to this event. In Section 1.3 we extend the applicability of the AIC to the model recently proposed by Watkins et al. (2014) that provides a simultaneous modeling of multiple event times, potentially right censored, in the presence of a non-susceptible group. While parametric survival models can be used as in the approach of Watkins et al. (2014), in this chapter we use the semiparametric Cox proportional hazard model for the susceptible part(s) of the mixture model and we use logistic regression for the incidence part. Simulation results are given in Section 1.4 and an application to credit loan data is presented in Section 1.5.

1.2 The mixture cure model for a single event

Mixture cure models were motivated by the existence of a subgroup of long-term survivors, or 'immunes' in a medical context. This subgroup, with survival probabilities set equal to one, is incorporated in a model through a mixture distribution where a logistic regression model provides a mixing proportion of the 'non-susceptible' cases and where a survival model describes the cases susceptible to the event of interest. Such models were introduced by Farewell (1982) in a parametric version, and later generalized to a semi-parametric mixture model combining logistic regression and Cox proportional hazards regression by Kuk and Chen (1992), see also Sy and Taylor (2000b). Recently, Cai et al. (2012a) introduced the R-package `smcure` to estimate such semi-parametric mixture models.

Tong et al. (2012) use a mixture cure approach to analyze the credit risk of a specific customer, where the event of interest is the time of default when customers stop paying back their loans. This setting is characterized and distinguishes itself from typical medical settings by a heavy right-censoring, since most customers do not default. A relatively large group of non-susceptible cases is expected to be present. Part of the explanation of this high percentage of censoring is that both prepayments and maturity (loan completely paid back on time) are considered censored for default. For a separate analysis of default and prepayment, see, e.g., Stepanova and Thomas (2002a).

1.2.1 Model notation

We denote the 'true' event time by U and the censoring time by C. We assume independence between event times and censoring times. Denote by Y a binary random variable where $Y = 1$ expresses susceptibility to the event of interest and $Y = 0$ indicates that the event will never happen. When $U > C$, the event is right-censored; the observed event time $T = \min(U, C)$. Let the indicator $\delta = I(U \leq C)$, thus $\delta = 1$ indicates non-censored observations. The combination of values for Y and δ generates three different states:

(1) $Y = 1$ and $\delta = 1$: uncensored and susceptible, so the event takes place during the observation period of the data;
(2) $Y = 1$ and $\delta = 0$: censored and susceptible, no event during the observation period, but it will eventually take place;
(3) $Y = 0$ and $\delta = 0$: censored and non-susceptible, no event is observed, nor will it take place in the future.

Note that values for T and δ are fully observed while Y is only observed when $\delta = 1$ and is latent otherwise. Similarly, we do not observe U when $\delta = 0$. The sample information consists of values (T_i, δ_i), for $i = 1, \ldots, n$, together with some covariate information.

The incidence model component uses logistic regression to model $P(Y = 1; \boldsymbol{z}) = \pi(\boldsymbol{z}, \boldsymbol{b})$ with $\text{logit}\{\pi(\boldsymbol{z}, \boldsymbol{b})\} = \boldsymbol{z}'\boldsymbol{b}$ for a r-vector of covariates $\boldsymbol{z} = (z_1, \ldots, z_r)'$. For the latency model, a semiparametric Cox proportional hazard regression model is used such that the survival probability at time t, conditional on $Y = 1$, is modeled as

$$S(t \mid Y = 1; \boldsymbol{x}, \boldsymbol{\beta}) \quad = \quad \exp\left(- \exp(\boldsymbol{x}^T \boldsymbol{\beta}) \int_0^t h_0(u \mid Y = 1) du \right),$$

with h_0 the unspecified baseline hazard function and $\boldsymbol{x}$ a q-vector of covariates $\boldsymbol{x} = (x_1, \ldots, x_q)'$, which may or may not contain the same components as $\boldsymbol{z}$. Denote that in our notation, while conditioning on Y, the arguments $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are separated by a semicolon, as these are the respective covariate and parameter vectors and no conditioning arguments. This yields the

so-called "unconditional" survival function, for given values of $\boldsymbol{x}, \boldsymbol{z}$ of the covariates and parameters $\boldsymbol{b}, \boldsymbol{\beta}$

$$S(t; \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{b}) = \pi(\boldsymbol{z}, \boldsymbol{b}) S(t \mid Y = 1; \boldsymbol{x}, \boldsymbol{\beta}) + 1 - \pi(\boldsymbol{z}, \boldsymbol{b}), \tag{1.1}$$

and the observed likelihood

$$\begin{aligned} L_{\text{obs}}(\boldsymbol{b}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \{\pi(\boldsymbol{z}_i, \boldsymbol{b}) f(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta})\}^{\delta_i} \\ &\quad \times \{(1 - \pi(\boldsymbol{z}_i, \boldsymbol{b})) + \pi(\boldsymbol{z}_i, \boldsymbol{b}) S(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta})\}^{1-\delta_i}, \end{aligned} \tag{1.2}$$

with $f(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta})$ the event density function for given covariate $\boldsymbol{x}$ and corresponding parameter vector $\boldsymbol{\beta}$, conditional on $Y = 1$. The relationship between this conditional event density function and the conditional survival function is given by

$$f(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta}) = h(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta})^{\delta_i} S(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta}),$$

with $h(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta})$ the conditional hazard function. The complete likelihood, given full information on Y, can be expressed as:

$$\begin{aligned} L_{\text{complete}}(\boldsymbol{b}, \boldsymbol{\beta}) &= \big(1 - \pi(\boldsymbol{z}_i, \boldsymbol{b})\big)^{(1-Y_i)} \pi(\boldsymbol{z}_i, \boldsymbol{b})^{Y_i} \\ &\quad \times h(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta})^{\delta_i, Y_i} S(t_i \mid Y_i = 1; \boldsymbol{x}_i, \boldsymbol{\beta})^{Y_i}. \end{aligned}$$

1.2.2 The Akaike information criterion for single event models

For estimation of mixture cure models, Cai et al. (2012a) explain the use of the expectation-maximization (EM) algorithm to deal with the latent Y values. If $Y = Y^*$ would be observed for all cases, the log-likelihood for the data triplets (T_i, δ_i, Y_i) could be used in the AIC to lead to the (infeasible)

$$\text{AIC}_{\text{infeasible}} = -2 \log L_{T,\delta,Y}(\widehat{\boldsymbol{\Theta}}; T_i, \delta_i, Y_i^*) + 2d, \tag{1.3}$$

where d counts the number of parameters in the model, and $\widehat{\boldsymbol{\Theta}}$ is the maximum likelihood estimator of the parameter vector $\boldsymbol{\Theta}$.

The AIC estimates the expected value of the Kullback-Leibler discrepancy between the model and the unknown true data-generating process, without having to know this true process.

In the general case with random variables $\boldsymbol{R} = (R_1, \ldots, R_n)$, a model $f(\boldsymbol{r}; \boldsymbol{\Theta})$, with $\boldsymbol{r}$ an instance of $\boldsymbol{R}$, and the density of the true data-generating process $g(\boldsymbol{r})$, the Kullback-Leibler discrepancy is given by $\mathrm{KL}\{g, f(\cdot; \boldsymbol{\Theta})\} = \mathrm{E}_g\Big\{ \log \dfrac{g(\boldsymbol{R})}{f(\boldsymbol{R}; \boldsymbol{\Theta})} \Big\}$, where the subscript g reminds of using the true density function g to compute the expectation. Define $\boldsymbol{\Theta}_0$ as the least false parameter value that minimizes the KL discrepancy between the model density $f(\cdot; \boldsymbol{\Theta})$ and the true density g, $\boldsymbol{\Theta}_0 = \arg\min_{\boldsymbol{\Theta}} \mathrm{KL}\{g, f(\cdot; \boldsymbol{\Theta})\}$. Since $\mathrm{E}_g[\log g(\boldsymbol{R})]$ does not vary when searching through several candidate models, minimizing $\mathrm{KL}\{g, f(\cdot; \boldsymbol{\Theta})\}$ over different models is equivalent with minimizing the quantity $D_{\boldsymbol{R}}(\boldsymbol{\Theta}) = \mathrm{E}_g\{-2 \log f(\boldsymbol{R}; \boldsymbol{\Theta})\}$, where the expectation is computed using the true density function of the data.

In our notation $\boldsymbol{R} = (T, \delta, Y)$, which can be split in an observed vector $\boldsymbol{R}_{\text{obs}} = (T, \delta)$ and a "missing" part $\boldsymbol{R}_{\text{mis}} = Y$ indicating that Y is not always observed. The expected complete-data log likelihood can be written as

$$\begin{aligned} Q^*(\boldsymbol{\Theta}) &= \mathrm{E}_{[Y]}[\log f_{T,\delta,Y}(T, \delta, Y; \boldsymbol{\Theta}) | T, \delta] \\ &= \sum_{i=1}^{n} \log f_{T,\delta,Y}(T_i, \delta_i, Y_i = 0; \boldsymbol{\Theta}) P(Y_i = 0 \mid T_i, \delta_i; \boldsymbol{\Theta}) \\ &\quad + \log f_{T,\delta,Y}(T_i, \delta_i, Y_i = 1; \boldsymbol{\Theta}) P(Y_i = 1 \mid T_i, \delta_i; \boldsymbol{\Theta}). \end{aligned}$$

Note that this is the conditional expectation of the log likelihood over Y given T and δ.

By rewriting the true joint density of the vector $\boldsymbol{R}$, with $r = (t, delta, y)$, as $g(r) = g_{Y|T,\delta}(y|t, \delta) \cdot g_{T,\delta}(t, \delta)$, and knowing that the expression of $Q^*(\boldsymbol{\Theta})$ is depending on T and δ, the expected value of $Q^*(\boldsymbol{\Theta})$ is $D_{\boldsymbol{R}}(\boldsymbol{\Theta}) = \mathrm{E}_{[T,\delta]}[-2Q^*(\boldsymbol{\Theta})]$. Because $\boldsymbol{\Theta}$ is estimated through $\widehat{\boldsymbol{\Theta}}$, $D_{\boldsymbol{R}}(\widehat{\boldsymbol{\Theta}})$ is a random variable and the AIC procedure estimates $\mathrm{E}[D_{\boldsymbol{R}}(\widehat{\boldsymbol{\Theta}})]$ using the sample information.

As used in the EM algorithm, for two values, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ of the param-

eter vector $\boldsymbol{\Theta} = (\boldsymbol{b}, \boldsymbol{\beta})$ the expected complete-data log likelihood applied to our problem can be estimated by, see also Cai et al. (2012a),

$$\begin{aligned} Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) &= \sum_{i=1}^{n} \log f_{T,\delta,Y}(T_i, \delta_i, Y_i{=}0; \boldsymbol{\Theta}_2) P(Y_i{=}0 \mid T_i, \delta_i; \boldsymbol{\Theta}_1) \\ &\quad + \log f_{T,\delta,Y}(T_i, \delta_i, Y_i{=}1; \boldsymbol{\Theta}_2) P(Y_i = 1 \mid T_i, \delta_i; \boldsymbol{\Theta}_1). \end{aligned} \tag{1.4}$$

Denote the first partial derivative $\dot{Q}(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) = \frac{\partial}{\partial \boldsymbol{\Theta_2}} Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1)$ and the second partial derivative $\ddot{Q}(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) = \frac{\partial}{\partial \boldsymbol{\Theta_2} \partial \boldsymbol{\Theta_2}'} Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1)$. The EM approach proceeds by maximizing $Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1)$ over $\boldsymbol{\Theta}_2$, and by replacing the current $\boldsymbol{\Theta}_1$ by the maximizer. These steps are iterated until convergence. The resulting value of $\boldsymbol{\Theta}$ is denoted by $\widehat{\boldsymbol{\Theta}}$.

In the context of missing data, Claeskens and Consentino (2008) prove in their Theorem 1 that for a model density f that is two times continuously differentiable with respect to $\boldsymbol{\Theta}$, and which has a bounded expectation of the second derivative in a neighborhood of $\boldsymbol{\Theta}_0$, which belongs to the interior of a compact parameter space, if $n(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)'(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)$ is uniformly integrable, with the prime denoting a transpose, then

$$\mathrm{E}[D_{\boldsymbol{R}}(\widehat{\boldsymbol{\Theta}}) - Q(\widehat{\boldsymbol{\Theta}} | \widehat{\boldsymbol{\Theta}})]/n = \mathrm{trace}\{\boldsymbol{I}^{-1}(\boldsymbol{\Theta}_0) \cdot \boldsymbol{J}(\boldsymbol{\Theta}_0)\}/n + o(1/n),$$

where $\boldsymbol{I}(\boldsymbol{\Theta}) = \mathrm{E}\{-\ddot{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta})/n\}$, and $\boldsymbol{J}(\boldsymbol{\Theta}) = \mathrm{Var}\{\dot{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta})\}/n$.

Following Cavanaugh and Shumway (1998), by first taking a Taylor series expansion of $\dot{Q}(\boldsymbol{\Theta}_0 \mid \widehat{\boldsymbol{\Theta}})$ around $\widehat{\boldsymbol{\Theta}}$, leads to estimate $\boldsymbol{J}(\boldsymbol{\Theta}_0)$ by $\boldsymbol{I}(\widehat{\boldsymbol{\Theta}}) \boldsymbol{I}_o^{-1}(\widehat{\boldsymbol{\Theta}}) \boldsymbol{I}(\widehat{\boldsymbol{\Theta}})$, and further to estimate $\boldsymbol{I}(\boldsymbol{\Theta}_0)$ by $\boldsymbol{I}_{oc}(\widehat{\boldsymbol{\Theta}})$, where

$$\boldsymbol{I}_{oc}(\widehat{\boldsymbol{\Theta}}) = -n^{-1} \frac{\partial^2 Q(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\Theta} \cdot \partial \boldsymbol{\Theta}'}, \qquad \boldsymbol{I}_o(\widehat{\boldsymbol{\Theta}}) = -n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \log f_{T,\delta}(T_i, \delta_i; \widehat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\Theta} \cdot \partial \boldsymbol{\Theta}'}.$$

This leads us to define the complete data AIC by

$$\mathrm{AICcd} = -2Q(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}) + 2 \ \mathrm{trace}\{\boldsymbol{I}_{oc}(\widehat{\boldsymbol{\Theta}}) \cdot \boldsymbol{I}_o^{-1}(\widehat{\boldsymbol{\Theta}})\}. \tag{1.5}$$

Note that this derivation has relaxed the strong assumption of Cavanaugh and Shumway (1998) to have the model correctly specified, that is, they assumed that $g(r) = f(r; \boldsymbol{\Theta}_0)$. By working with least false parameter values, we avoided this strong assumption.

The computation of $\boldsymbol{I}_o$, which requires the joint density of (T, δ), not including Y, is facilitated by the use of the supplemented EM-algorithm (Meng and Rubin, 1991). The EM-algorithm implicitly defines a mapping $\boldsymbol{\Theta} \to \boldsymbol{M}(\boldsymbol{\Theta}) = (M_1(\boldsymbol{\Theta}), \ldots, M_d(\boldsymbol{\Theta}))'$ from the parameter space to itself such that $\widehat{\boldsymbol{\Theta}}^{(m+1)} = \boldsymbol{M}(\widehat{\boldsymbol{\Theta}}^{(m)})$ for $m = 0, 1, \ldots$. A Taylor series expansion in the neighborhood of $\widehat{\boldsymbol{\Theta}}$ yields that

$$(\widehat{\boldsymbol{\Theta}}^{(m+1)} - \widehat{\boldsymbol{\Theta}})' \approx (\widehat{\boldsymbol{\Theta}}^{(m)} - \widehat{\boldsymbol{\Theta}})' \boldsymbol{DM}, \text{ where } \boldsymbol{DM} = \left(\frac{\partial M_j(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}_i} \right) \Bigg|_{\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}},$$

a $d \times d$-matrix evaluated at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$. Meng and Rubin (1991) further show that $\boldsymbol{I}_o^{-1} = \boldsymbol{I}_{oc}^{-1}(\boldsymbol{I}_d - \boldsymbol{DM})^{-1}$, with $\boldsymbol{I}_d$ a $d \times d$ identity matrix. For more details on the computation of $\boldsymbol{DM}$, we refer to Chap. 12 of Gelman et al. (2004) and Section 3.3 of Meng and Rubin (1991). Using (1.5), this leads to the AICcd as we use it in this chapter,

$$\begin{aligned} \text{AICcd} &= -2Q(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}) + 2\,\text{trace}(\boldsymbol{I}_d - \boldsymbol{DM})^{-1} \\ &= -2Q(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}) + 2d + 2\,\text{trace}\{\boldsymbol{DM}(\boldsymbol{I}_d - \boldsymbol{DM})^{-1}\}. \quad (1.6) \end{aligned}$$

This criterion differs in two aspects from the infeasible AIC in (1.3). First, the expected complete data likelihood is used, and second, there is a correction to the penalty term that takes the complexity of the modeling process due to the missing information into account. When all data are observed, $\boldsymbol{DM} = 0$ and the penalty reduces to the classical one.

We wish to mention that the mixture regression criterion of Naik et al. (2007) as an extension of the AIC to select both the number of components in the mixture and the variables within each component is not suitable for our purpose. Indeed, we know exactly the number of components in the mixture from the problem content, moreover even partial cluster membership is known. Only for censored observations the group membership is unknown. In addition, the mixture regression criterion assumes fully observed cases, while these data here are intrinsically censored.

1.2.3 AIC explicitly incorporating censoring

An alternative treatment of the censored observations is to treat the censored times as "missing" event times. The model that we wish to find should be well for describing the true event times U, and not only for the observed times T. Therefore, we start by writing the joint log likelihood of (U, Y) as, with $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{b})$,

$$L_n(\boldsymbol{\Theta}; U, Y) = \sum_{i=1}^{n} \Big\{ \log P_Y(\boldsymbol{z}_i, \boldsymbol{b}) + \log \tilde{f}_Y(U_i; \boldsymbol{\beta}) \Big\},$$

where $P_Y(\boldsymbol{z}_i) = \pi(\boldsymbol{z}_i, \boldsymbol{b})$ when $Y_i = 1$ and $P_Y(\boldsymbol{z}_i) = 1 - \pi(\boldsymbol{z}_i, \boldsymbol{b})$ when $Y_i = 0$. Note that, with C_i the censoring time for observation i, if $T_i \leq C_i$, the true event time is observed and $U_i = T_i$, while if $T_i > C_i$, the true event time U_i is not observed.We define $\tilde{f}_Y(u_i; \boldsymbol{\Theta}) = f_{T|Y}(t_i \mid Y_i = 1; \boldsymbol{\beta})^{\delta_i} f_{U|Y}(u_i \mid Y_i = 1; \boldsymbol{\beta})^{(1-\delta_i)}$ when $Y_i = 1$ and take $\tilde{f}_Y(u_i) = 1$ when $Y_i = 0$. The Q-function for use in the EM-algorithm and the AIC can here be defined as,

$$\begin{aligned} Q(\boldsymbol{\Theta}_2; \boldsymbol{\Theta}_1) \;=\; & \sum_{i=1}^{n} \Big(\log \pi(\boldsymbol{z}_i, \boldsymbol{b}_2) \\ & + \mathrm{E}_{[U]}\big[\log \big\{ \tilde{f}_Y(U_i \mid Y_i = 1; \boldsymbol{\Theta}_2) \big\} \mid T_i; \boldsymbol{\Theta}_1 \big] \Big) w_{1i}(\boldsymbol{\Theta}_1) \\ & + \sum_{i=1}^{n} \Big(\log\, (1 - \pi(\boldsymbol{z}_i, \boldsymbol{b}_2)) + \log \big\{ 1 \big\} \Big) (1 - w_{1i}(\boldsymbol{\Theta}_1)), \end{aligned}$$

where $w_{1i}(\boldsymbol{\Theta}_1) = P(Y_i = 1 \mid T_i = t; \boldsymbol{\Theta}_1)$ and the expectation '$\mathrm{E}_{[U]}$' is here computed with respect the model density of true event times U, given $Y = 1$ and using parameter value $\boldsymbol{\Theta}_1$. Recall that, if $T_i \leq C_i$, the true event time is observed and $U_i = T_i$. Then we have that

$$\begin{aligned} & \mathrm{E}_{[U]}\big[\log \big\{ \tilde{f}_Y(U_i \mid Y_i = 1; \boldsymbol{\Theta}_2) \big\} \mid T_i; \boldsymbol{\Theta}_1 \big] \\ = \; & \sum_{i=1}^{n} \delta_i \log f_{T|Y}(T_i \mid Y_i = 1; \boldsymbol{\Theta}_2) \\ & + \sum_{i=1}^{n} (1 - \delta_i) \mathrm{E}_{[U]}[\log f_{U|Y}(U_i \mid Y_i = 1; \boldsymbol{\Theta}_2) \mid T_i; \boldsymbol{\Theta}_1]. \end{aligned}$$

This leads to defining the function Q for use in an EM-algorithm in the following way,

$$\begin{aligned} Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) = & \sum_{i=1}^{n} \log \pi(\boldsymbol{z}_i, \boldsymbol{b}_2) w_{1i}(\boldsymbol{b}_1, \boldsymbol{\beta}_1) \\ + & \sum_{i=1}^{n} \log(1 - \pi(\boldsymbol{z}_i, \boldsymbol{b}_2))\{1 - w_{1i}(\boldsymbol{b}_1, \boldsymbol{\beta}_1)\} \\ + & \sum_{i=1}^{n} \delta_i \log f_{T|Y}(T_i \mid Y_i = 1; \boldsymbol{\beta}_2) w_{1i}(\boldsymbol{b}_1, \boldsymbol{\beta}_1) \\ + & \sum_{i=1}^{n} (1 - \delta_i) \frac{\int_{c_i}^{\infty} \log f_{U|Y}(u_i \mid Y_i = 1; \boldsymbol{\beta}_2) f_{U|Y}(u_i \mid Y_i = 1; \boldsymbol{\beta}_1) du_i}{P(T_i \geq C_i \mid Y_i = 1; \boldsymbol{\beta}_1)} w_{1i}(\boldsymbol{b}_1, \boldsymbol{\beta}_1), \end{aligned}$$

with

$$\begin{aligned} w_{1i}(\boldsymbol{\Theta}) &= P(Y_i = 1 \mid T_i = t; \boldsymbol{\Theta}) \\ &= \begin{cases} \dfrac{\pi(\boldsymbol{z}_i; \boldsymbol{b}) \log f_{U|Y}(U_i \mid Y_i = 1; \boldsymbol{\beta})}{\pi(\boldsymbol{z}_i, \boldsymbol{b}) \log f_{U|Y}(U_i \mid Y_i = 1; \boldsymbol{\beta})) + (1 - \pi(\boldsymbol{z}_i, \boldsymbol{b}))} & \text{for } \delta_i = 0 \\ 1 & \text{for } \delta_i = 1. \end{cases} \end{aligned}$$

Defining the AICcd proceeds as in Section 1.2.2 using this function Q. The resulting AICcd has a correct Kullback-Leibler interpretation for right-censored data from a mixture distribution. This way of incorporating the censoring provides (in models without mixture) an alternative to the AIC proposed by Suzukawa et al. (2001).

1.3 AIC for multiple event mixture cure models

We extend the parametric competing risk model of Watkins et al. (2014) by allowing for the semiparametric Cox proportional hazard model. In this model one distinguishes multiple events (e.g., default, prepayment) for which the time to event is important and considers another class of events (such as maturity) which happen at a fixed time. This class encompasses the group of 'immunes' in Section 1.2. For the multiple event mixture cure model, censored loans are the loans that are still being repaid. As a result,

although these loans will eventually experience one of these three events, the eventual outcome is not clear yet. For the formulation of this model, three indicators are used:

(1) Y_m, indicating that the loan is considered to be mature, so repayed at the indicated end date of the loan;
(2) Y_d, indicating that default takes place;
(3) Y_p, indicating that early repayment takes place.

Note that this set of (Y_m, Y_d, Y_p) is exhaustive and mutually exclusive. However, when an observation is censored, it is not known which event type will occur. In analogy to the eqs. (1.1) and (1.2), the survival function, unconditional on the Y-triplet for given values of the covariates $\boldsymbol{x}_p, \boldsymbol{x}_d$ and $\boldsymbol{z}$, can be written as (denote $\pi_p(\boldsymbol{z}) = P(Y_p = 1; \boldsymbol{z})$, $\pi_d(\boldsymbol{z}) = P(Y_d = 1; \boldsymbol{z})$)

$$\begin{aligned} S(t; \boldsymbol{x}_p, \boldsymbol{x}_d, \boldsymbol{z}) &= \pi_p(\boldsymbol{z}) S_p(t \mid Y_p = 1; \boldsymbol{x}_p) \\ &\quad + \pi_d(\boldsymbol{z}) S_d(t \mid Y_d = 1; \boldsymbol{x}_d) + \big(1 - \pi_p(\boldsymbol{z}) - \pi_d(\boldsymbol{z})\big), \end{aligned}$$

with S_p and S_d denoting the survival functions for, respectively, prepayment and default. Note that mature loans are handled as special cases, as maturity should not be considered as a real event. Hence, there is no survival function for maturity, or the survival function could simply be considered to be equal to one until the maturity date. Using the subscript '1' for default (d) and '2' for prepayment (p), the corresponding observed likelihood is given by

$$\begin{aligned} L_{\text{obs}}(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \Big\{ \prod_{j=1}^{2} \big(\pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j) f_j(t_i \,|\, Y_{j,i} = 1; \boldsymbol{x}_{j,i}, \boldsymbol{\beta}_j)\big)^{Y_{j,i}} \big(1 - \sum_{j=1}^{2} \pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j)\big)^{Y_{m,i}} \Big\}^{\delta_i} \\ \times \Big\{ \big(1 - \sum_{j=1}^{2} \pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j)\big) + \sum_{j=1}^{2} \pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j) S_j(t_i \,|\, Y_{j,i} = 1; \boldsymbol{x}_{j,i}, \boldsymbol{\beta}_j) \Big\}^{1-\delta_i}, \end{aligned}$$

where $\boldsymbol{\Theta} = (\boldsymbol{b}_p, \boldsymbol{b}_d, \boldsymbol{\beta}_p, \boldsymbol{\beta}_d)$. Note the flexibility of this model; each model part may employ its own set of covariates, hence the vectors $\boldsymbol{x}_d, \boldsymbol{x}_p$ and $\boldsymbol{z}$ may be different. We rewrite this model for use in an EM algorithm such that the AICcd of (1.6) may be applied for model selection. For this purpose, we start from the complete likelihood, hence the likelihood expression

under the assumption that full information on $\boldsymbol{Y} = (Y_m, Y_d, Y_p)$ is present

$$
\begin{aligned}
&L_{\text{complete}}(\boldsymbol{\Theta}; \delta_i, Y_i, T_i) \\
&\quad = \prod_{i=1}^{n} \Big\{ \prod_{j=1}^{2} \big(\pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j)\big)^{Y_{j,i}} \big(1 - \prod_{j=1}^{2} \pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j)\big)^{Y_{m,i}} \Big\} \\
&\quad \times \Big\{ \prod_{j=1}^{2} \big(h_j(t \mid Y_{j,i} = 1; \boldsymbol{x}_{j,i}, \boldsymbol{\beta}_j)^{\delta_i} S_j(t_{j,i} \mid Y_{j,i} = 1; \boldsymbol{x}_{j,i}, \boldsymbol{\beta}_j)\big)^{Y_{j,i}} \Big\}.
\end{aligned}
$$

Converting to the log likelihood and computing the expected value this time using the model density with parameter $\boldsymbol{\Theta}_1$ leads us to the Q-function as given in (1.4),

$$
\begin{aligned}
Q(\boldsymbol{\Theta}_2 \mid \boldsymbol{\Theta}_1) &= \mathrm{E}_f[\log L_{\text{complete}}(\boldsymbol{\Theta_2}; T_i, \delta_i, Y_i) \mid T_i, \delta_i; \boldsymbol{\Theta_1}] \\
&= \sum_{i=1}^{n} \Big\{ \sum_{j=1}^{2} w_{ji} \log(\pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j)) + w_{mi} \log(1 - \sum_{j=1}^{2} \pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j)) \\
&\qquad + \sum_{j=1}^{2} \delta_i \log(h_j(t_i \mid Y_j = 1; \boldsymbol{x}_{j,i}, \boldsymbol{\beta}_j)) \\
&\qquad + w_{ji} \log(S_j(t_i \mid Y_j = 1; \boldsymbol{x}_{j,i}, \boldsymbol{\beta}_j)) \Big\}.
\end{aligned}
$$

Note that conditional expectations of $Y_{j,i}$ $(j = 1, 2)$, $\mathrm{E}_f\big[Y_{j,i} \mid T_i, \delta_i; \boldsymbol{\Theta}_1\big]$, are computed here with respect to the model density using parameter $\boldsymbol{\Theta}_1$ and are denoted by w_{ji} with $w_{mi} = 1 - w_{1i} - w_{2i}$ and for $j = 1, 2$,

$$
\begin{aligned}
w_{ji} &= w_{ji}(\boldsymbol{\Theta}) = P(Y_j = 1 \mid T = t_i, \delta_i; \boldsymbol{\Theta}) \\
&= \begin{cases}
\dfrac{\pi_j(\boldsymbol{z}_i, \boldsymbol{b}_j) S_j(t_i; \boldsymbol{\beta}_j)}{\sum_{k=1}^{2} \pi_k(\boldsymbol{z}_i, \boldsymbol{b}_k) S_k(t_i; \boldsymbol{\beta}_k) + (1 - \sum_{k=1}^{2} \pi_k(\boldsymbol{z}_i, \boldsymbol{b}_k))} & \text{for } \delta_i = 0 \\
1 & \text{for } Y_{j,i} = 1;\ \delta_i = 1 \\
0 & \text{for } Y_{j,i} = 0;\ \delta_i = 1.
\end{cases}
\end{aligned}
$$

1.4 Simulation study

1.4.1 Simulation settings

All computations were performed in R (R Core Team, 2014), adapting the library `smcure` (Cai et al., 2012a) to produce the AICcd values.

variable	v_1	v_2	v_3	v_4	v_5
Distr.	Ber(0.7)	$\Gamma(\lambda = 2.74, r = 1.3086)$	$N(1,1)$	$N(1,2)$	Ber(0.66)

Table 1.1: *Distributions of z_1–z_5 used in the simulation study.*

Three different simulation settings were used. For each simulation setting, 100 simulation runs with n=5000 observations and 5 variables were executed. The probability of being susceptible, that is $(1 - \pi(\boldsymbol{v}))$ was generated using the relationship $\pi(\boldsymbol{v}) = \dfrac{\exp(\boldsymbol{b}'\boldsymbol{v})}{1 + \exp(\boldsymbol{b}'\boldsymbol{v})}$, with variables v_1–v_5 of which the distributions are stated in Table 1.1 and with parameters $\boldsymbol{b}$ as in Table 1.2. True Y-values are consequently generated via a Bernoulli distribution using these probabilities $\pi(\boldsymbol{v})$. For the uncured part of the population, Weibull default times (shape parameter = 1, scale parameter = 0.5) were generated, using the same five variables v_1–v_5 with the distributions and parameter values $\boldsymbol{\beta}$ as in Tables 1.1 and 1.2. For the first two simulation settings, censoring times were uniformly distributed on the interval $[0, 1]$. For setting 3, censoring times were uniformly distributed on the interval $[0, 20]$, in order to lower the amount of censoring compared to settings 1 and 2. Each time we performed an exhaustive model search, thus $(2^5 - 1)^2 = 961$ AICcd's were calculated for every simulation run. The purpose of this simulation study is to examine to what extent the AICcd is capable of selecting the correct covariate vectors $\boldsymbol{x} = (v_1, v_4)$ and $\boldsymbol{z} = (v_1, v_2, v_5)$. Note that, because of the presence of all types of covariate distributions in the credit risk context, the distributions in Table 1.1 are not restricted to normal distributions, but also gamma and Bernoulli distributions are included.

In the first simulation setting, the censoring percentage was 60% (hence, around 3000 observations were censored, $\delta = 0$) and 80% of the observations were susceptible ($Y = 1$). For setting 2, we mimicked the situation of the data example (see Section 1.5), resulting in the uncensored percentage nearly equal to 10%, and the susceptible percentage of the observations equal to 20%. For setting 3 the censoring time interval was increased

parameter	(intercept)	b_1	b_2	b_3	b_4	b_5	β_1	β_2	β_3	β_4	β_5
Setting 1 & 3	3	3.5	-1	0	0	-1	2.5	0	0	-1	0
Setting 2	1	1.5	-1.5	0	0	-1.8	2.5	0	0	-1	0

Table 1.2: *Simulation study. Parameter values of the true model.*

from [0,1] to [0,20], resulting in more observed defaults, and less censoring. Because of this, the real default time was observed for 70% of the observations, with 80% susceptible observations as in setting 1.

For comparison purposes, four other versions of AIC were calculated:

$$\begin{aligned}
\mathrm{AIC}_{\mathrm{cs}} &= -2\log L_{\mathrm{Cox}}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{x}\right) + 2d_{\mathrm{Cox}},\\
\mathrm{AIC}_{\mathrm{cl}} &= -2\log L_{\mathrm{Cox}}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{x}\right) + 2d_{\mathrm{Cox,Log}},\\
\mathrm{AIC}_{\mathrm{ls}} &= -2\log L_{\mathrm{Log}}\left(\hat{\boldsymbol{b}}, \boldsymbol{z}\right) + 2d_{\mathrm{Log}},\\
\mathrm{AIC}_{\mathrm{ll}} &= -2\log L_{\mathrm{Log}}\left(\hat{\boldsymbol{b}}, \boldsymbol{z}\right) + 2d_{\mathrm{Cox,Log}}.
\end{aligned}$$

The first subscript of the AIC's is either c or l, which stands for "Cox" or "Log" and indicates the likelihood of the survival or logistic part of the mixture only. The second subscript indicates whether a "short" (s) or "long" (l) penalty term was used. A short penalty term means that the parameters accounted for are only calculated by the model specified in the first subscript, and a long penalty term incorporates all the parameters. The penalty is defined to be twice the number of considered parameters.

The reason for comparing the AICcd to those at first sight rather naive AIC-calculations, is because in practice, those AICs might by some researchers be in use instead of the corrected version with complete-data log likelihoods when analyzing mixture cure models. We want to investigate whether it is reasonable to use those AICs. We are not aware of other model selection criteria for mixture cure models.

Settings	Method	Mean rank	Log -	Log +	Cox -	Cox +	Total -	Total +
MAX		961	3	2	2	3	5	5
1	AICcd	107.85	1.14	0.90	0.02	1.68	1.16	2.58
	AICcs	163.13	1.63	1.51	0.00	0.66	1.63	2.17
	AICcl	163.91	1.70	1.19	0.00	0.67	1.70	1.86
	AICls	155.26	1.43	0.44	0.67	1.58	2.10	2.02
	AICll	151.67	1.46	0.48	0.67	1.13	2.13	1.61
2	AICcd	59.81	0.00	1.32	0.17	1.44	0.17	2.76
	AICcs	95.06	0.88	1.51	0.01	0.83	0.89	2.34
	AICcl	94.95	1.02	1.07	0.01	0.76	1.03	1.83
	AICls	162.64	0.02	1.46	1.99	1.55	2.01	3.01
	AICll	159.58	0.02	1.43	2.00	1.17	2.02	2.60
3	AICcd	13.01	0.00	0.84	0.00	0.91	0.00	1.75
	AICcs	79.53	1.28	1.49	0.00	0.41	1.28	1.90
	AICcl	80.39	1.35	0.99	0.00	0.41	1.35	1.40
	AICls	151.68	2.58	1.16	1.06	2.32	3.64	3.48
	AICll	147.33	2.59	1.17	1.06	2.02	3.65	3.19

Table 1.3: *Simulation settings 1 – 3, 100 runs for an exhaustive search. Averages for underfitting (-) and overfitting (+) in terms of variables as compared to the true model, for each part of the mixture model, and for the combined parts (total).*

1.4.2 Simulation results

Table 1.3 summarizes some model selection aspects of the AICs. The results of all simulation runs were averaged. Next to the type of AIC used, we list the ranking (among the 961 models) of the true model as simulated. The next four columns indicate the average number of variables that were lacking in the selected model (-) or were unnecessarily included in the selected model (+) for the log-component and the Cox-component respectively as compared to the "true" model. The last two columns are the joint averaged over- and underselection values.

The simulated data were generated using three true variables for the log-model, and two variables for the Cox-model. The first line in Table 1.3 indicates the maximum value possible for each column of the table. A

perfect selection would give a mean rank of 1 (= the "true" model is always selected), and 0-values for all the other entries, indicating that all necessary variables are present in the model, and all the unnecessary variables are left out. AIC is known to be an efficient model selection method with regard to mean squared prediction error (Claeskens and Hjort, 2008, Chap 4), though not to be consistent hence we do not expect to find small ranks for the true model here. Moreover, the chosen settings are quite demanding with large percentages of censored data (especially for settings 1 and 2), which are typical to credit risk studies, as opposed to medical studies where those percentages are usually much smaller.

The simulation study indicates that for these settings the Cox part of the log-likelihood is dominant, both in magnitude and for model selection purposes. In Table 1.3, we see that AICcd outperforms the other criteria regarding the mean rank of the true model for all three settings. Overfitting proportions are favorable for the low-censored setting (setting 3), but quite high for setting 1 and 2. On the other hand, underfitting proportions are low for the AICcd compared to the other measures. This is an important result as underfitting (missing important predictors) is considered worse than overfitting. When looking at the change in result as the censoring percentage changes, it becomes clear that high percentages of censored cases on one hand (setting 2) and a big discrepancy between observed versus true defaults (setting 1) have a negative impact on the performance of any information criterion. This gives us a strong indication that it would be advisable to incorporate additional information (such as in the multiple event models) to reduce the number of censored cases.

A comparison with the simpler criterion that just counts the number of parameters is for the chosen settings not behaving too badly, since it turns out that the correction term involving $\boldsymbol{DM}$ takes values in a bounded range, and is here not influencing the model order too much. Again, we stress that no other information criteria have yet been developed for these mixture models, which could have made the comparison more interesting. For comparisons of AICcd in regression models to other AIC-like versions we refer to Cavanaugh and Shumway (1998).

	Description	Type
v_1	The gender of the customer (1=M, 0=F)	Categorical
v_2	Amount of the loan	Continuous
v_3	Number of years at current address	Continuous
v_4	Number of years at current employer	Continuous
v_5	Amount of insurance premium	Continuous
v_6	Home phone or not (1=N,0=Y)	Categorical
v_7	Own house or not (1=N, 0=Y)	Categorical
v_8	Frequency of payment(1=low/unknown, 0=high)	Categorical

Table 1.4: *Credit loan data. Description of the variables.*

1.5 Variable selection for a credit loan dataset

1.5.1 Data and method

The survival analysis techniques were applied to personal loan data from a major UK financial institution. All customers are UK borrowers who had applied to the bank for a loan. The data set consisted of the application information of 50,000 personal loans, together with the repayment status for each month of the observation period of 36 months. We note that the same data were also used in Stepanova and Thomas (2002a) and later by Tong et al. (2012). In this chapter only a subset of the loans with loan term 36 months were used for the analysis (containing n =7521 observations).

An account was considered as a default (censoring indicator=1) if it was at least 90 days in arrears. When an account was not in arrears or only in arrears for less than 90 days, the account was considered as a non-default (censoring indicator=0). As for most credit data, the percentage of defaults within the observation period was very low: default was only observed for 376 of the 7521 observations. In Section 1.5.3 we reconsider this dataset taking prepayments and maturity into account, hereby reducing the number of censored cases.

For each observation, we considered eight candidate covariates, see Ta-

ble 1.4. In the model selection approach of Section 1.5.2, we searched through all subsets of the collection of eight covariates, and this for both model components, resulting in $(2^8 - 1)^2 = 65025$ AICcd values, where we have excluded empty latency and incidence models. Using the same method of exhaustive search for the modeling approach in Section 1.5.3 would result in over 16 581 375 AICcd calculations ($(2^8 - 1)^3$), because this time three different covariate vectors are considered. Therefore, instead of an exhaustive search, a genetic algorithm was used to find a good model, for which we used the package `GA` in R (Scrucca, 2013). We used this package with AICcd in the binary representation indicating the presence (1) or absence (0) of a specific variable, and with all default settings, i.e., population size 50, crossover probability 0.8, mutation probability 0.1. For the model selection purpose, this algorithm starts with randomly including and excluding some variables. The algorithm consists of several "generations", and at the end of each generation, the AICcd-values of the inspected models are evaluated, and the models with the lowest AICcd-values are withheld in the next generations. Starting from those models, small changes are made with the purpose to find models with even lower AICcd-values.

1.5.2 Variable selection for the time to default

After calculating the AICcd values for each of the considered models, the models were sorted according to their resulting AICcd values. Seven models will be discussed and compared: the five best models according to the AICcd, the full model and (again according to AICcd) the best model under the restriction that the latency and incidence model should contain the same covariates; see Table 1.5.

We observe that for all the five best models, the same latency model is selected whereas the incidence model covariates vary. For this dataset, the incidence model seems to require more variables. Whereas variables v_2 (amount of the loan), v_3 (number of years living at a current address) and v_8 (frequency of the payment) are never included in the latency part

Model	AICcd	Rank	Part	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Best	7372.85	1	Incidence	1	0	1	1	1	1	1	1
			Latency	1	0	0	1	1	1	1	0
Second best	7373.06	2	Incidence	1	1	1	1	1	1	1	1
			Latency	1	0	0	1	1	1	1	0
Third best	7385.11	3	Incidence	1	1	0	1	1	1	1	1
			Latency	1	0	0	1	1	1	1	0
Fourth best	7385.28	4	Incidence	1	1	1	1	1	1	1	0
			Latency	1	0	0	1	1	1	1	0
Fifth best	7385.79	5	Incidence	1	0	1	1	1	1	1	0
			Latency	1	0	0	1	1	1	1	0
Full	7446.92	215	(Both)	1	1	1	1	1	1	1	1
Same covariate	7397.87	17	(Both)	1	0	0	1	1	1	1	0

Table 1.5: *Credit loan data. Variables contained in the five best models according to AICcd, the full model and the AICcd-best model with the same parameters in both model parts. The value of AICcd, as well as its ranking is given.*

of the best five models, those three variables are also the ones left out in the incidence model, but at the most with two at the same time. The full model only ranks 215th with regard to AICcd value. The same covariate model, for this dataset, uses the same covariates as for the latency part of the best five models. Its rank is 17, with a difference in AICcd values as compared to the best model equal to about 25, clearly showing the preference for the separate covariate parts.

In the credit risk context, a widely used method to evaluate binary classifiers is by means of the receiver operating characteristics curves. These curves give the percentage of correctly classified observations for each possible threshold value. The specific measure of interest is the area under the curve (AUC), which can also be used in the context of survival analysis (Heagerty and Saha, 2000). We computed the AUC values for five models of interest, when predicting default at three different time instances (18,

Month	Best	Second best	Third best	Full	Same covariate
18	0.710	0.709	0.695	0.707	0.703
24	0.700	0.700	0.683	0.700	0.688
36	0.688	0.685	0.664	0.684	0.671

Table 1.6: *Credit loan data. AUC values for the top three models according to AICcd, the full model and the AICcd-best model with the same variables in both model parts, when predicting default at 18, 24 and 36 months respectively.*

Part	Int.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Inc. ($\hat{\boldsymbol{b}}$)	-1.586	-0.311	–	-0.036	-0.044	0.001	0.002	0.328	-0.380
(se)	(0.210)	(0.155)	–	(0.009)	(0.014)	(0.0002)	(0.285)	(0.129)	(0.120)
Lat. ($\hat{\boldsymbol{\beta}}$)	–	0.551	–	–	-0.066	0.0003	0.852	0.024	–
(se)	–	(0.177)	–	–	(0.019)	(0.0002)	(0.304)	(0.172)	–

Table 1.7: *Credit loan data. The parameter estimates for the time to default with their standard errors (se) for the AICcd-best model for the incidence (Inc.) and latency (Lat.) parts of the model. Variables not selected were not estimated.*

24 and 36 months). Each time, 2/3 of the data was used as a training set, and 1/3 as a test set. The AUC-values can be found in Table 1.6.

In Table 1.7, the parameter estimates of the best model according to AICcd can be found. Positive $\boldsymbol{b}$-parameters have a positive impact on the probability of being susceptible, and positive $\boldsymbol{\beta}$-parameters shorten the time until default. As a result, working at the same employer for a longer time period decreases the risk to default, as well as having a home phone and owning a house (binary variables decoded as 1 = no and 2 = yes). The gender of a subject has an ambiguous effect on default: whereas being male lowers the probability of being susceptible, we see that the time until default when susceptible is shorter for men.

Figure 1.1 presents the estimated survival curves for two randomly

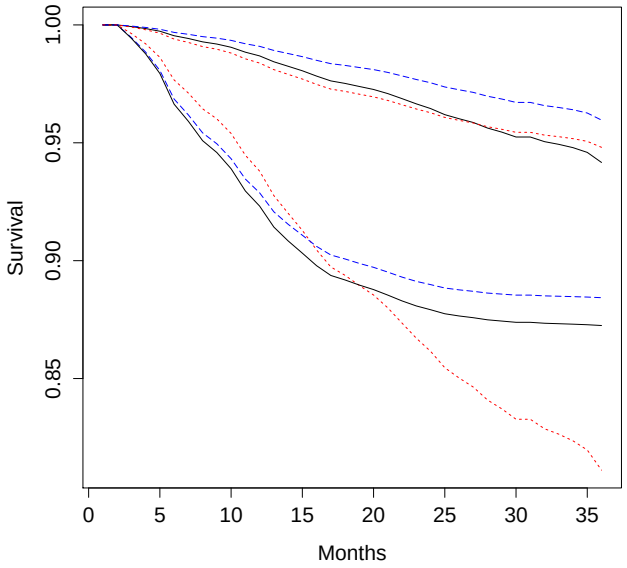

Figure 1.1: *Credit loan data. Estimated survival curves for two observations using three models. In solid line type (black) we show the estimates for the selected best model, the dashed lines (blue) use the same-covariate best model, while the dotted lines (red) give the estimated survival curve using the Cox proportional hazard model, ignoring the mixture.*

chosen persons in the dataset (namely a male person, not possessing a home phone and working at the same employer for a relatively short time, and a female person, possessing a home phone and working at the same employer for a relatively long time). We consider estimates obtained in the best mixture cure model with different covariates for both model parts, in the best such model with the same covariates, and in the best Cox proportional hazard model with all variables except for the customer's gender. This was the model selected by the AIC using the partial likelihood and penalizing for the number of parameters in the model.

For the female person, the estimated survival percentages were relatively high, and all three approaches give reasonably close estimates.

However, for the male person with lower values for the estimated survival probabilities, we observe a clear difference with the estimates from the mixture model and with that of the Cox proportional hazard model. The estimated proportion in the mixture was equal to 12.81 % for this subject, clearly suggesting the need of the mixture model. For this data example, the use of the same covariates leads to larger estimated probabilities for survival.

1.5.3 Variable selection for the multiple event model

As stated before, the multiple event model does not only incorporate default, but also early repayment, resulting in two incidence models and two latency models. For this dataset there are 3.6% observations (269 cases) for which maturity has occurred (so, which are belonging to the "cured" fraction), 5% (376 cases) were in default, and 39.8% (2992 cases) have prepayments. The remaining 51.6% are truly censored observations.

The genetic algorithm used is part of the package `GA` in R by Scrucca (2013), with default settings, as described in Section 1.5.1. Despite the fact that genetic algorithms are quite successful and efficient, it is never certain that the final outcome will yield the overall lowest AICcd value. However, the genetic algorithm we used was also applied to the data example for the mixture cure model in Section 1.5.2, resulting in precisely the same selected model as with the exhaustive search. The resulting model for the joint analysis of default and prepayment with parameter estimates can be found in Table 1.8. The interpretation of the parameters in Table 1.8 is similar to the mixture cure-interpretation. Again, we see that not having a home phone increases the probability and shortens the time for default (both positive $\hat{\boldsymbol{b}}$- and $\hat{\boldsymbol{\beta}}$ values). A longer working duration at the same employer, however, decreases the probability of default but has no significant result on the time until default according to the model selected by the genetic algorithm using AICcd. The number of parameters included in the latency model of default has gone from five parameters in the mixture cure model to four parameters in the multiple event incidence model. A

Part	Intercept	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
$\hat{\boldsymbol{b}}_d$	-0.837	-0.084	-0.00007	-0.038	-0.094	0.001	0.481	–	-0.479
$\hat{\boldsymbol{\beta}}_d$	–	0.118	–	–	–	0.0001	0.342	–	0.106
$\hat{\boldsymbol{b}}_p$	0.648	-0.174	-0.00001	-0.020	-0.014	-0.00001	0.083	–	-0.084
$\hat{\boldsymbol{\beta}}_p$	–	0.073	-0.00003	–	–	–	-0.359	-0.081	0.163

Table 1.8: *Credit loan data. The parameter estimates for the multiple event incidence model as found by the genetic algorithm.*

possible explanation is that since more information is gained by adding an early repayment part, less predictors are needed for the time until default. For the early repayment parameters, we notice that five variables are included in the latency part. We see that male subjects tend to have a lower chance to belong to the early repayment group ($\boldsymbol{b} < 0$), but when belonging to that group, they tend to prepay earlier than female subjects. Note that the same variables are included for the two incidence models, where only v_7 is not in the incidence model. This is a result of the fact that the respective probabilities are estimated in one multinomial logit model (as we have now three groups: early repayment, default and maturity). The sign of $\hat{\boldsymbol{b}}_{\boldsymbol{d}}$ and $\hat{\boldsymbol{b}}_{\boldsymbol{p}}$ gives the relation between default and early repayment respectively, in relation to maturity. For example: the multinomial log-odds for a certain subject to belong to the early repayment-group versus the mature group are expected to increase by 0.083 units (ceteris paribus) when the subject does not have a homephone, however, the log-odds to belong to the default-group compared to the mature group are even more elevated (increase by 0.481 units).

As a final illustration, the default and early repayment curves were plotted in Figure 1.2 for the same two random observations as for the mixture cure model. The male person incurs a higher risk regarding default, and a lower propensity regarding early repayment.

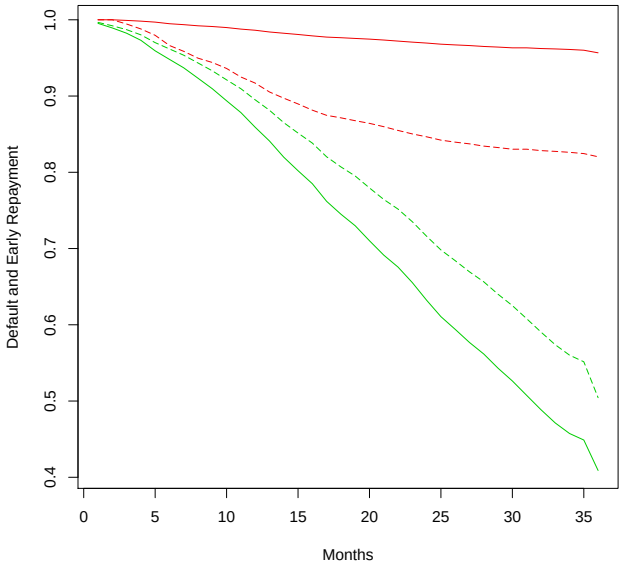

Figure 1.2: *Credit loan data. Estimated probabilities for default and early repayment for two observations. The green (steeper) lines represent early repayment, and the flatter lines default. The solid line represents a female person, possessing a home phone and working at the same employer for a relatively long time, and the dashed lines a male person, not possessing a home phone and working at the same employer for a relatively short time.*

1.6 Discussion

The development of advanced survival models for credit risk data is in current progress. With this chapter we contributed with the derivation of a proper variable selection method. We have used the popular Akaike information criterion as the basis of the selection procedure. By making use of the output of the EM procedure for model fitting, we obtained a relatively simple criterion and have implemented this procedure in R, making use of existing packages for fitting mixture cure models.

Our simulation study and the data analysis have illustrated that using different covariate vectors may lead to better models regarding AUC value and regarding to model ranking according to AICcd. Not restricting to same-covariate models for mixture modeling is worthwhile, our variable selection approach easily allows for such general modeling strategies. The use of a genetic search algorithm in combination with the AICcd provides a handy way of incorporating many variables.

Chapter 2

A hierarchical mixture cure model with unobserved heterogeneity using the EM-algorithm

Abstract

The specific nature of credit loan data requires the use of mixture cure models within the class of survival analysis tools. The constructed models allow for competing risks such as early repayment and default on one hand, and for incorporating maturity, expressed as an insusceptible part of the population, on the other hand. This chapter further extends such models by incorporating unobserved heterogeneity within the risk groups. A hierarchical expectation-maximization algorithm is derived to fit the models and standard errors are obtained. Simulations and a data analysis illustrate the applicability and benefits of these models, and in particular an improved event time estimation.

This chapter is based on Dirick, L., Claeskens, G., Vasvnev, A. and Baesens, B. (2015). A hierarchical mixture cure model with unobserved heterogeneity using the EM-algorithm. Working paper, submitted.

2.1 Introduction

The analysis of credit risks via survival analysis takes advantage of the nature of time-to-event data, in particular by its ability to naturally capture the specifics of default, prepayment and maturity events. While first those events were examined and modeled individually, see for example Banasik et al. (1999), Stepanova and Thomas (2002a), Andreeva (2006) and Bellotti and Crook (2009), these models were soon extended by allowing for a cured fraction while modeling early repayment or default, known as mixture cure models, see Tong et al. (2012) and Dirick et al. (2015). The simultaneous analysis of all different events is evident in Deng et al. (2000), Pavlov (2001), Ciochetti et al. (2002), Dirick et al. (2015) and Watkins et al. (2014).

In this chapter we extend such multiple event models for credit risk data by acknowledging the fact that there are different kinds of customers. For example, some people are risk-averse, others might be risk-neutral or even risk-seeking. While this characteristic is not observed, our approach makes it possible to incorporate unobserved heterogeneity in the models. More specifically, we construct an expectation-maximization (EM) algorithm that simultaneously deals with the mixture cure model with multiple events and with a number of subgroups for each of the modeled events. We explain the implementation of such a hierarchical EM algorithm for the credit risk models.

This chapter gives the first simulation study of the mixture cure models with unobserved heterogeneity. An application of the model for personal loan data from a UK bank reiterates the importance of the unobserved heterogeneity for credit risks. In the simultaneous modeling of competing events, similar to Watkins et al. (2014) we find that the explanatory variables can act in different directions upon incidence and duration; and, variables exist that are statistically significant in explaining only incidence or duration.

In another context, Deng et al. (2000) showed that there exists significant heterogeneity among mortgage borrowers which generated discussion

in this area, though not many researchers followed their lead. Recently Burda et al. (2015) employed an approach to build a semiparametric competing risk model with unobserved heterogeneity for the analysis of unemployment in the US. Their Bayesian method does not involve the EM-algorithm, and introduces unobserved heterogeneity through an infinite mixture of generalized inverse Gaussian densities.

Despite the fact that in this chapter the focus is on competing risks for loan data, the model is not restricted to these types of data and is applicable in a large range of situations where competing risks (and a possibility of not undergoing the risk or an "insusceptible" part of the population) and a certain amount of censoring are present. In the biomedical context, many disease-related research uses these models when there are several possible death causes (for example Lunn and McNeil, 1995), when there is a cured fraction of the patients (see, e.g., Bremhorst and Lambert, 2014) or the combination of both (e.g. Ng et al., 2002). In the economic context, an interesting example is given in Berrington and Diamond (2000), where first-partnership formation (competing risks are cohabitation and marriage) of males and females born in 1958 in Britain is studied. An insusceptible population part can then be defined as the subjects that will never marry or cohabitate, however, censoring is present through the subjects that have not yet entered the first-partnership at the moment of the study, but will afterwards. Burda et al. (2015) use another application where the time to moment of exit from unemployment is modeled (to the same versus another industry where they had been employed previously).

The chapter is organized as follows. Section 2.2 gives the hierarchical mixture cure model with unobserved heterogeneity. Section 2.3 details the EM-algorithm. The simulation study is summarized in Section 2.4 and the empirical application is in Section 2.5. Concluding comments are in Section 2.6 followed by the theoretical derivations in the appendix.

2.2 The hierarchical mixture cure model

We observe life times T_i and a set of covariates. The life times T_i represent the time until an event $j \in \{1, \ldots, J\}$ takes place, or until the observation is censored. In the latter case, the general censoring indicator δ_i for observation i is equal to 0, indicating that none of the competing events was observed. Additionally, each observation has J event-specific censoring indicators, denoted by $\delta_{j,i}$. As it is assumed that events are mutually exclusive, the rationale is that the occurrence of a certain event causes the observation to be censored from any other event type. Note that $\delta_i = \sum_{y=1}^{J} \delta_{y,i}$. For censored observations ($\delta_{j,i} = 0$ for every j and, consequently, $\delta_i = 0$), it is unknown which of the event types will be experienced eventually, or in other words, the event "group" that a censored observation belongs to is unknown. This group membership is represented by a partially observed variable $Y \in \{1, \ldots, J\}$, with Y being observed only when $\delta_i = 1$.

Denote by $\pi_j(\boldsymbol{z}, \boldsymbol{b}) = P(Y = j; \boldsymbol{z}, \boldsymbol{b})$ the probability of belonging to a certain group j, with $j \in \{1, \ldots, J\}$, given the covariate vector $\boldsymbol{z}$ and the vector of coefficients $\boldsymbol{b}$. For this discrete distribution it holds that (for a fixed z) $0 \leq \pi_j(\boldsymbol{z}, \boldsymbol{b}) \leq 1$ and that $\sum_{j=1}^{J} \pi_j(\boldsymbol{z}, \boldsymbol{b}) = 1$. The estimation of $\pi_j(\boldsymbol{z}, \boldsymbol{b})$ is done through a multinomial logistic regression model with a covariate vector z and corresponding parameter vector b. For $j = J$ it holds that $\pi_J(\boldsymbol{z}, \boldsymbol{b}) = 1 - \sum_{y=1}^{J-1} P(Y = y; \boldsymbol{z}, \boldsymbol{b})$ with for $j \in \{1, \ldots, J-1\}$,

$$\pi_j(\boldsymbol{z}, \boldsymbol{b}) = P(Y = j; \boldsymbol{z}, \boldsymbol{b}) = \frac{\exp(\boldsymbol{z}^T \boldsymbol{b}_j)}{1 + \sum_{y=1}^{J-1} \exp(\boldsymbol{z}^T \boldsymbol{b}_y)}. \tag{2.1}$$

The probability of not having experienced any event by time t is then given by

$$S(t; \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{b}, \boldsymbol{\beta}) = \sum_{y=1}^{J-1} \pi_y(\boldsymbol{z}, \boldsymbol{b}) S(t \mid Y = y; \boldsymbol{x}, \boldsymbol{\beta}_y) + \pi_J(\boldsymbol{z}, \boldsymbol{b}), \tag{2.2}$$

where $S(t \mid Y = j; \boldsymbol{x}, \boldsymbol{\beta}_j)$ is the probability of not having experienced the event j by time t. The proof of the identifiability of (2.2) is in Appendix

2.7.1. The survival probabilities $S(t \mid Y = j; \boldsymbol{x}, \boldsymbol{\beta}_j)$ use a covariate vector $\boldsymbol{x}$, which may be different from, overlapping with, or be identical to the covariate vector $\boldsymbol{z}$.

In group J the subjects are insusceptible to any of the considered events, or in other words "cured", which is originating from medical studies considering cured patients, see e.g. Kuk and Chen (1992), Sy and Taylor (2000b), Peng and Dear (2000). In the model, the cured or insusceptible group has a survival probability $S(t \mid Y = J) = 1$ for every t and does not depend on x or on any parameters.

To incorporate heterogeneity, in a hierarchical model we assume that all $J-1$ main groups, thus except for the "insusceptible" Jth group, may be further divided into K_j subgroups, of which observations experience the same event and have a similar covariate structure but differ with regard to their event time structure. So instead of immediately modeling survival function $S(t \mid Y = j, \boldsymbol{x}; \boldsymbol{\beta}_j)$ which depends on main group membership only, the survival structure depends on the subgroups as well. The probability of not having experienced event j at time t when belonging to subgroup k is modeled by a semi-parametric Cox proportional hazards model and given by

$$\begin{aligned} & S_{T|\widetilde{Y}_j,Y}(t \mid \widetilde{Y}_j = k, Y = j; \boldsymbol{x}, \boldsymbol{\beta}_{jk}) \\ &= \exp\Big\{ - \exp(\boldsymbol{x}^T \boldsymbol{\beta}_{jk}) \int_0^t h_0(u \mid \widetilde{Y}_j = k, Y = j) du \Big\}, \end{aligned} \tag{2.3}$$

with h_0 the unspecified baseline hazard function, estimated using Breslow's estimator. The latent variable $\widetilde{Y}_j$ which takes values in $\{1, \ldots, K_j\}$ represents the subgroup membership for group $j \in \{1, \ldots, J-1\}$ and $\boldsymbol{\beta}_{jk}$ is the parameter vector related to the survival function of subgroup k in main group j. For further use we define the probability of belonging to subgroup k as $\tau_{k|j} = P(\widetilde{Y}_j = k \mid Y = j)$.

2.3 The joint likelihood and EM-algorithm for the hierarchical model

The likelihood contribution of an observation $i = 1, \ldots, n$ is given by

$$\begin{aligned} & f_{T,\delta,\widetilde{Y},Y}(t_i, \delta_i, \tilde{y}, y; \boldsymbol{b}, \boldsymbol{\beta}_{Y,\widetilde{Y}}) \\ &= f_{T,\delta|\widetilde{Y},Y}(t_i, \delta_i \mid \widetilde{Y}_i = \tilde{y}, Y_i = y; \boldsymbol{x}_i, \boldsymbol{\beta}_{y\tilde{y}}) \cdot \tau_{\widetilde{Y}_i|Y_i} \cdot \pi_{Y_i}(\boldsymbol{z}_i, \boldsymbol{b}), \end{aligned}$$

where, for a given j and k, the likelihood contribution is, recall $\delta_i = \sum_{y=1}^{J} \delta_{y,i}$:

$$\begin{aligned} & f_{T,\delta|\widetilde{Y}_j,Y}(t_i, \delta_i \mid \widetilde{Y}_j = k, Y = j; \boldsymbol{x_i}, \boldsymbol{\beta}_{jk}) \\ & \quad = h_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk})^{\delta_{j,i}} \\ & \quad \times S_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk}), \end{aligned} \tag{2.4}$$

with h being the hazard function, formally,

$$h_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk}) = h_0(t \mid \widetilde{Y}_j = k, Y = j) \exp(\boldsymbol{x}^T \boldsymbol{\beta}_{jk}).$$

The joint hierarchical log likelihood of $(T_i, Y_i, \widetilde{Y}_i)$ now takes the form

$$\begin{aligned} & \mathcal{L}_n^h(\boldsymbol{b}, \boldsymbol{\beta}_{Y,\widetilde{Y}}; Y_1, \ldots, Y_n, \widetilde{Y}_1, \ldots, \widetilde{Y}_n, T_1, \ldots, T_n, \delta_1, \ldots, \delta_n) \\ &= \log\Big(\prod_{i=1}^{n} f_{T,\delta,\widetilde{Y},Y}(t_i, \delta_i, \tilde{y}, y; \boldsymbol{b}, \boldsymbol{\beta}_{y\tilde{y}})\Big) \\ &= \log\Big(\prod_{i=1}^{n} \big\{ f_{T,\delta|\widetilde{Y},Y}(T_i, \delta_i \mid \widetilde{Y}_i, Y_i; \boldsymbol{x}_i, \boldsymbol{\beta}_{Y,\widetilde{Y}}) \cdot \tau_{\widetilde{Y}_i|Y_i} \cdot \pi_{Y_i}(\boldsymbol{z}_i, \boldsymbol{b}) \big\}\Big). \end{aligned} \tag{2.5}$$

2.3.1 The expected complete-data log likelihood

Since the main group indicators Y_i as well as the subgroup indicators $\widetilde{Y}_i$ are not fully observed, the EM-algorithm (Dempster et al., 1977) is used in order to maximize the log likelihood. In this iterative procedure, the parameter estimates of the r-th EM-iteration are used along with the observed information to compute the expected complete-data log likelihood of the $(r+1)$-th EM-iteration, formally,

$$
\begin{aligned}
&Q^h\big\{(\boldsymbol{b},\boldsymbol{\beta}_{Y,\widetilde{Y}})^{(r+1)} \mid (\boldsymbol{b},\boldsymbol{\beta}_{Y,\widetilde{Y}})^{(r)}\big\} \\
&= \sum_{i=1}^{n} E[\log f_{T,\delta,\widetilde{Y},Y}(T_i,\delta_i,\widetilde{Y}_i,Y_i;\boldsymbol{b},(\boldsymbol{\beta}_{Y,\widetilde{Y}})^{(r+1)}) \mid T_i;(\boldsymbol{\beta}_{Y_i,\widetilde{Y}_i})^{(r)}] \\
&= \sum_{i=1}^{n} \Big\{ E[\log \pi_{Y_i}(\boldsymbol{z}_i,\boldsymbol{b}^{(r+1)}) \mid T_i;\boldsymbol{b}^{(r)}] + E[\log \tau^{(r+1)}_{\widetilde{Y}_i|Y_i} \mid T_i;\boldsymbol{\beta}^{(r)}_{Y_i,\widetilde{Y}_i}] \\
&\qquad + E[\log f_{T|Y,\widetilde{Y}}(T_i,\delta_i \mid Y_i,\widetilde{Y}_i;\boldsymbol{x}_i,\boldsymbol{\beta}^{(r+1)}_{Y,\widetilde{Y}}) \mid T_i;\boldsymbol{\beta}^{(r)}_{Y_i,\widetilde{Y}_i}] \Big\}. \qquad (2.6)
\end{aligned}
$$

Note that this is the expected value of the likelihood conditional on T_i, with parameter estimates of iteration r. Rewriting the first term gives

$$
E[\log \pi_{Y_i}(z_i,\boldsymbol{b}^{(r+1)}) \mid T_i;\boldsymbol{b}^{(r)}] = \sum_{y=1}^{J} P(Y_i = y \mid T_i = t_i,\boldsymbol{x}_i) \log \pi_y(\boldsymbol{z_i},\boldsymbol{b}^{(r+1)}),
$$

with $P(Y_i = j \mid T_i = t_i;\boldsymbol{x}_i)$ the probability of belonging to group j, conditional on the censoring or event time. This probability is for uncensored cases either 1 or 0. It is a weighted average of the time densities in the censored case,

$$
\begin{aligned}
&P(Y_i = j \mid T_i = t_i,\boldsymbol{x}_i) \\
&= \begin{cases}
\dfrac{\pi_j(\boldsymbol{z}_i,\boldsymbol{b}^{(r)}) f_{T|Y=j}(t_i \mid Y = j;\boldsymbol{x}_i,\boldsymbol{\beta}^{(r)}_j)}{\sum_{y=1}^{J} \pi_y(\boldsymbol{z}_i,\boldsymbol{b}^{(r)}) f_{T|Y=y}(t_i \mid Y = y;\boldsymbol{x}_i,\boldsymbol{\beta}^{(r)}_y)} & \text{if } \delta_i = 0 \\
1 \qquad \text{if } \delta_i = 1 \text{ and } Y_i = j \\
0 \qquad \text{if } \delta_i = 1 \text{ and } Y_i \neq j
\end{cases} \\
&\equiv w_j(\boldsymbol{\beta}^{(r)};t_i,\boldsymbol{x}_i).
\end{aligned}
$$

As the density $f_{T|Y=j}(t_i \mid Y = j;\boldsymbol{x}_i,\boldsymbol{\beta}^{(r)}_j)$ itself is composed of the subgroup time densities with their respective subgroup probabilities, $\tau_{j|k}$,

$$
\begin{aligned}
&w_j(\boldsymbol{\beta}^{(r)};t_i,\boldsymbol{x}_i) \\
&= \begin{cases}
\dfrac{\pi_j(\boldsymbol{z}_i,\boldsymbol{b}^{(r)}) \sum_{\tilde{y}=1}^{K_j} \tau_{\tilde{y}|j} f_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = \tilde{y}, Y = j;\boldsymbol{x}_i,\boldsymbol{\beta}^{(r)}_{j\tilde{y}})}{\sum_{y=1}^{J} \pi_y(\boldsymbol{z}_i,\boldsymbol{b}^{(r)}) \sum_{\tilde{y}=1}^{K_y} \tau_{\tilde{y}|y} f_{T|\widetilde{Y}_y,Y}(t_i \mid \widetilde{Y}_y = \tilde{y}, Y = y;\boldsymbol{x}_i,\boldsymbol{\beta}^{(r)}_{y\tilde{y}})} & \text{if } \delta_i = 0 \\
1 \qquad \text{if } \delta_i = 1 \text{ and } Y_i = j \\
0 \qquad \text{if } \delta_i = 1 \text{ and } Y_i \neq j,
\end{cases}
\end{aligned}
$$

and using (2.4) along with the fact that $\delta_{j,i} = 0$ when $\delta_i = 0$,

$$w_j(\boldsymbol{\beta}^{(r)}; t_i, x_i)$$
$$= \begin{cases} \dfrac{\pi_j(\boldsymbol{z}_i, \boldsymbol{b}^{(r)}) \sum_{\tilde{y}=1}^{K_j} \tau_{\tilde{y}|j} S_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = \tilde{y}, Y = j; \boldsymbol{x}_i, \boldsymbol{\beta}_{j\tilde{y}}^{(r)})}{\sum_{y=1}^{J} \pi_y(\boldsymbol{z}_i, \boldsymbol{b}^{(r)}) \sum_{\tilde{y}=1}^{K_y} \tau_{\tilde{y}|y} S_{T|\widetilde{Y}_y,Y}(t_i \mid \widetilde{Y}_y = \tilde{y}, Y = y; \boldsymbol{x}_i, \boldsymbol{\beta}_{y\tilde{y}}^{(r)})} & \text{if } \delta_i = 0 \\ 1 \quad \text{if } \delta_i = 1 \text{ and } Y_i = j \\ 0 \quad \text{if } \delta_i = 1 \text{ and } Y_i \neq j. \end{cases} \tag{2.7}$$

For the second term in (2.6) we get that

$$E[\log \tau_{\widetilde{Y}_i|Y_i}^{(r+1)} \mid T_i; \boldsymbol{\beta}_{Y_i,\widetilde{Y}_i}^{(r)}] = \sum_{y=1}^{J} \sum_{\tilde{y}=1}^{K_y} P(\widetilde{Y}_{j,i} = \tilde{y}, Y_i = y \mid T_i = t_i, \boldsymbol{x}_i) \log \tau_{\tilde{y}|y}^{(r+1)}$$
$$= \sum_{y=1}^{J} \sum_{\tilde{y}=1}^{K_y} P(\widetilde{Y}_{j,i} = \tilde{y} \mid Y_i = y; T_i = t_i, \boldsymbol{x}_i) P(Y_i = y \mid T_i = t_i, \boldsymbol{x}_i) \log \tau_{\tilde{y}|y}^{(r+1)},$$

with $P(\widetilde{Y}_{j,i} = k \mid Y_i = j; T_i = t_i, \boldsymbol{x}_i)$ the probability of belonging to subgroup k, given the event type j and the censoring or event time. Similarly to (2.7), we get

$$\begin{aligned} & P(\widetilde{Y}_{j,i} = k \mid Y_i = j, T_i = t_i; \boldsymbol{x}_i) \\ &= \frac{\tau_{k|j}^{(r)} f_{T|Y=j,\widetilde{Y}_j=k}(t_i \mid Y = j, \widetilde{Y}_j = k; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk}^{(r)})}{\sum_{\tilde{y}=1}^{K_j} \tau_{\tilde{y}|j}^{(r)} f_{T|Y=j,\widetilde{Y}_j=\tilde{y}}(t_i \mid Y = j, \widetilde{Y}_j = \tilde{y}; \boldsymbol{x}_i, \boldsymbol{\beta}_{j\tilde{y}}^{(r)})} \\ &\equiv v_{k|j}(\beta_j^{(r)}; t_i, \boldsymbol{x}_i), \end{aligned}$$

for $j \in \{1, \ldots, J-1\}$ and $k \in \{1, \ldots, K_j\}$. By consequence,

$$E[\log \tau_{\widetilde{Y}_i|Y_i}^{(r+1)} \mid T_i; \boldsymbol{\beta}_{Y_i,\widetilde{Y}_i}^{(r)}] = \sum_{y=1}^{J-1} \sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|y}(\boldsymbol{\beta}_y^{(r)}; t_i, \boldsymbol{x}_i) w_y(\boldsymbol{\beta}^{(r)}; t_i, \boldsymbol{x}_i) \log \tau_{\tilde{y}|y}^{(r+1)}.$$

As opposed to the main groups where $\delta_{j,i}$ gives partial information on membership, no prior information is available for subgroup membership thus τ does not depend on a covariate vector. In the first iteration of the EM-algorithm, a value for $\tau_{k|j}$ is chosen for each k such that $\sum_{k=1}^{K} \tau_{k|j} = 1$ for each j. Without prior information, a logical starting value for $\tau_{k|j}$ is $1/K_j$ with K_j the total number of subgroups in main group j. In the next

steps of the EM-algorithm, $\tau_{k|j}$ is updated as follows (see Appendix 2.7.2 for details),

$$\tau_{k|j}^{(r+1)} = \tau_{k|j}^{(r+1)}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = P(\widetilde{Y}_j = k \mid Y = j) = \frac{\sum_{i=1}^{n} v_{k|j_i}^{(r)}(\boldsymbol{\beta}_j^{(r)}, t_i, \boldsymbol{x}_i)}{n}. \quad (2.8)$$

Similarly, the third term of (2.6) is given by

$$\begin{aligned} & E[\log f_{T_i|Y_i,\widetilde{Y}_i}(T_i \mid Y_i, \widetilde{Y}_i; \boldsymbol{x}_i, \boldsymbol{\beta}_{Y,\widetilde{Y}}^{(r+1)}) \mid T_i; \boldsymbol{\beta}_{Y,\widetilde{Y}}^{(r)}] \\ &= \sum_{y=1}^{J}\sum_{\tilde{y}=1}^{K_y} v_{\tilde{y}|y}(\boldsymbol{\beta}_y^{(r)}; t_i, \boldsymbol{x}_i) w_y(\boldsymbol{\beta}^{(r)}; t_i, \boldsymbol{x}_i) \\ &\quad \times \log f_{T_i,y_i,\tilde{y}_i}(T_i \mid Y_i = y, \widetilde{Y}_i = \tilde{y}; \boldsymbol{x}_i, \boldsymbol{\beta}_{y\tilde{y}}^{(r+1)}) \\ &= \sum_{y=1}^{J-1}\sum_{\tilde{y}=1}^{K_y} v_{\tilde{y}|y}(\boldsymbol{\beta}_y^{(r)}, t_i, \boldsymbol{x}_i) w_y(\boldsymbol{\beta}^{(r)}; t_i, \boldsymbol{x}_i) \\ &\quad \times \log f_{T_i,y_i,\tilde{y}_i}(T_i \mid Y_i = y, \widetilde{Y}_i = \tilde{y}; \boldsymbol{x}_i, \boldsymbol{\beta}_{y\tilde{y}}^{(r+1)}). \end{aligned}$$

The resulting hierarchical complete-data log likelihood is then given by

$$\begin{aligned} & Q^h\{(\boldsymbol{b}, \boldsymbol{\beta}_{Y,\widetilde{Y}})^{(r+1)} \mid (\boldsymbol{b}, \boldsymbol{\beta}_{Y,\widetilde{Y}})^{(r)}\} \quad (2.9) \\ &= \sum_{i=1}^{n}\Bigg[w_J(\boldsymbol{\beta}^{(r)}; t_i, \boldsymbol{x}_i) \log \pi_J(\boldsymbol{z}_i, \boldsymbol{b}^{(r+1)}) \\ &\quad + \sum_{y=1}^{J-1}\Big\{ w_y(\boldsymbol{\beta}^{(r)}; t_i, \boldsymbol{x}_i) \log \pi_y(\boldsymbol{z}_i, \boldsymbol{b}^{(r+1)}) \\ &\quad + \sum_{\tilde{y}=1}^{K_y} w_y(\boldsymbol{\beta}^{(r)}; t_i, \boldsymbol{x}_i) v_{\tilde{y}|y}(\boldsymbol{\beta}_y^{(r)}; t_i, \boldsymbol{x}_i)\Big[\log \tau_{\tilde{y}|y}^{(r+1)} \\ &\quad + \log f_{T_i,y_i,\tilde{y}_i}(T_i \mid Y_i = y, \widetilde{Y}_i = \tilde{y}; \boldsymbol{x}_i, \boldsymbol{\beta}_{y\tilde{y}}^{(r+1)})\Big]\Big\}\Bigg], \end{aligned}$$

2.3.2 Initialization and iterative E- and M-step

The three main steps of the computational algorithm are performed as follows:

a) Initialization stage

1) Determine the *number of subgroups* K_j for each of the J-1 main groups. Whereas the number of main groups J is fixed by the data structure, the number of subgroups is not. It is suggested to try several values for K_j, and evaluate the results (see also Section 2.5.2).

2) *Initialization of w:* Set $w_j^{(0)}(\boldsymbol{\beta}^{(0)}; t_i, \boldsymbol{x}_i) = \delta_{j,i}$ for every j. Hence, the initial value is 1 for an observed event of category j and is 0 otherwise.

3) *Initialization of* $\boldsymbol{b}$*:* Fit a multinomial logit model to $w^{(0)}$ using covariate vector $\boldsymbol{z}$, in order to retrieve an initial estimate $\hat{\boldsymbol{b}}^{(0)}$.

4) *Initialization of* $\boldsymbol{\beta}$*:* Obtain estimates $\hat{\boldsymbol{\beta}}_{j,k}^{(0)}$ at each subgroup level. The parameter estimates of the multiple event mixture cure model (Dirick et al., 2015) without heterogeneity can be used to set the initial values for the $\sum_{y=1}^{J-1} K_y$ parameter vectors. *Remark:* The K_j initial values for every $j \in \{1, \ldots, J-1\}$ should be different for the algorithm to work more efficiently.

5) *Initialization of* τ*:* $\tau_{k|j}^{(0)} = 1/K_j$ if no information about subgroups.

6) *Initialization of densities:* Compute density $f_{T|Y=j,\tilde{Y}_j=k}(t_i \mid Y = j, \widetilde{Y}_j = k; \boldsymbol{x}_i, \boldsymbol{\beta}_{j,k})$, baseline hazard $h_0(u \mid \widetilde{Y}_j = k, Y = j)$ through Breslow's estimator and the survival function $S_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j; \boldsymbol{x}_i, \boldsymbol{\beta}_{j,k})$ values (using (2.3)) for each observation, using the initial $\hat{\boldsymbol{\beta}}_{j,k}^{(0)}$-estimates.

b) E-step

1) Compute $\pi_j^{(1)}(\boldsymbol{z}_i, \boldsymbol{b})$ for each j, using $\hat{\boldsymbol{b}}^{(0)}$.

2) Compute $w_j^{(1)}(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i)$ for each j, using $\hat{\boldsymbol{\beta}}^{(0)}$.

3) Compute $v_{k|j}^{(1)}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i)$ for each k and each j, using $\hat{\boldsymbol{\beta}}_j^{(0)}$.

c) M-step

1) *Update* $\boldsymbol{b}$*:* Obtain a new estimate $\hat{\boldsymbol{b}}^{(1)}$ using the $w_j^{(1)}(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i)$'s of the E-step.

2) *Update $\boldsymbol{\beta}_{jk}$:* Obtain a new estimate $\hat{\boldsymbol{\beta}}_{j,k}^{(1)}$ using mixture weights $w_j^{(1)}(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i)$ and $v_{k|j}^{(1)}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i)$. *Method:* The likelihood contribution corresponding to the event times can be written as

$$\begin{aligned}
&\sum_{j=1}^{J-1}\sum_{k=1}^{K_j}\sum_{i=1}^{n} w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i) v_{k;j}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i) \\
&\qquad \times \log\Big\{ h_{T,j,k}(t_i; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk})^{\delta_{j,i}} S_{T,j,k}(t_i; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk}) \Big\} \\
= &\sum_{j=1}^{J-1}\sum_{k=1}^{K_j}\sum_{i=1}^{n} w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i) v_{k;j}(\boldsymbol{\beta}_j, t_i, \boldsymbol{x}_i) \delta_{j,i} \log h_{T,j,k}(t_i; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk}) \\
&+ w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i) v_{k;j}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i) \log S_{T,j,k}(t_i; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk}) \\
= &\sum_{j=1}^{J-1}\sum_{k=1}^{K_j}\sum_{i=1}^{n} v_{k;j}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i) \Big(\delta_{j,i} \log\big\{ w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i) h_{T,j,k}(t_i; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk}) \big\} \\
&+ w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i) \log S_{T,j,k}(t_i; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk}) \Big).
\end{aligned}$$

For the last step, we used $\log w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i)\delta_{j,i} = 0$ and $\delta_{j,i} w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i) = \delta_{j,i}$. we can use $\log(w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i))$ as an offset variable.
This way of writing illustrates that $\boldsymbol{\beta}$ can be estimated using standard software for fitting Cox proportional hazards models, such as the `coxph`-function in `R`, with an additional offset variable $\log(w_j(\boldsymbol{\beta}; t_i, \boldsymbol{x}_i))$ and weights equal to $v_{k|j}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i)$. A similar reasoning has been used by Cai et al. (2012a).

3) *Update densities:* Obtain a new estimate of $f_{T|Y=j,\widetilde{Y}_j=k}(t_i \,|\, Y = j, \widetilde{Y}_j = k; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk})$, $h_0(u \mid \widetilde{Y}_j = k, Y = j)$ and $S_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk})$, using $\hat{\boldsymbol{\beta}}_{j,k}^{(1)}$.

Repeat the E- and M-step with all updated estimates, until parameter convergence. The algorithm stops when the sum of the absolute value of the relative differences between $\big(\hat{\boldsymbol{\beta}}^{(r+1)}, \hat{\boldsymbol{b}}^{(r+1)}\big)$ and $\big(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\boldsymbol{b}}^{(r)}\big)$ is smaller than 10^{-6}.

Note that a computational issue can arise for some observations, as $f_{T|Y=j,\widetilde{Y}_j=k}(t_i \mid Y = j, \widetilde{Y}_j = k; \boldsymbol{x}_i, \boldsymbol{\beta}_{jk})$ for all subgroups $k \in \{1, \ldots, K_j\}$

in one of the main groups $j \in \{1, \ldots, (J-1)\}$ can be very small or even 0. As a result, $v_{k|j}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i)$ can go to infinity for all K_j subgroups of group j. As $\sum_k^{K_j} v_{k|j}$ should be equal to 1, this issue is solved by putting $v_{k|j}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i) = 1/K_j$ for every k when the denominator of $v_{k|j}(\boldsymbol{\beta}_j; t_i, \boldsymbol{x}_i) < 10^{-10}$ for a certain observation i. Intuitively, when a certain observation is seemingly found to not belong to any of the subgroups of a main group j, its subgroup probabilities should be equal.

2.3.3 Standard errors through the SEM-algorithm

The typical execution of the EM-algorithm does not automatically produce standard errors of the parameter estimates, for an overview, see Jamshidian and Jennrich (2000). This method, introduced by Meng and Rubin (1991), is widely used in various applications for standard error estimation when applying the EM-algorithm, see for example Segal et al. (1994) and Cai and Lee (2009). Meng and Rubin (1991) show that a numerically stable asymptotic variance-covariance matrix can be obtained using the supplemented EM-algorithm, more specifically, $V = I_{oc}^{-1}(I_d - \boldsymbol{DM})^{-1}$, where I_{oc} is the negative second hessian matrix of the expected complete-data log likelihood, with $\boldsymbol{\Theta} = (\boldsymbol{b}, \boldsymbol{\beta}_{Y,\widetilde{Y}})$, define

$$I_{oc} = -\frac{\partial^2 Q^h\big(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}\big)}{\partial \boldsymbol{\Theta} \cdot \partial \boldsymbol{\Theta}^T}.$$

The matrix I_d is the identity matrix with dimension $d \times d$, with d equal to the length of the parameter vector Θ. The $d \times d$-matrix DM can be interpreted as the matrix rate of convergence of the EM-algorithm. The idea behind this is that, implicitly, a mapping $\Theta \to M(\boldsymbol{\Theta}) = (M_1(\boldsymbol{\Theta}), \ldots, M_d(\boldsymbol{\Theta}))^T$ is defined by the EM-algorithm from the parameter space to itself such that for $r = 0, 1, \ldots$, $M(\hat{\boldsymbol{\Theta}}^{(r)}) = \hat{\boldsymbol{\Theta}}^{(r+1)}$. A Taylor series expansion in the neighborhood of $\hat{\boldsymbol{\Theta}}$ yields that

$$(\widehat{\boldsymbol{\Theta}}^{(r+1)} - \widehat{\boldsymbol{\Theta}})' \approx (\widehat{\boldsymbol{\Theta}}^{(r)} - \widehat{\boldsymbol{\Theta}})' \boldsymbol{DM}, \text{ where } \boldsymbol{DM} = \left(\frac{\partial M_l(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta_m}}\right)\Bigg|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}},$$

a $d \times d$-matrix evaluated at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$. In practice, $\boldsymbol{DM}$ is obtained by numerically differentiating $M(\boldsymbol{\Theta})$. For more information on the computa-

tion of $\boldsymbol{DM}$, we refer to Meng and Rubin (1991, Section 3.3). We have implemented this procedure to obtain standard errors of the estimators.

2.4 Simulation study

2.4.1 Simulation settings

To validate the model, a simulation study was conducted. In each simulation run, a dataset of size $n = 7500$ was constructed with three event groups A, B and C, and two subgroups for groups A and B (C is the "cured" group). There are three variables, generated respectively by $x_1 \sim \mathcal{N}(-2, 1)$, $x_2 \sim \mathcal{N}(2.6, 1.2)$ and $x_3 \sim \mathcal{N}(3, 2)$. These variables in combination with $\boldsymbol{b}$-parameters shape the event group memberships of the observations when using a multinomial logit transformation. As a result, we get the probability of belonging to group $j \in \{A, B\}$:

$$\pi_{i,j}(\boldsymbol{z}_i, \boldsymbol{b}) = P(Y = j; \boldsymbol{b}) = \frac{\exp(\boldsymbol{z}_i^T \boldsymbol{b}_j)}{1 + \exp(\boldsymbol{z}_i^T \boldsymbol{b}_A) + \exp(\boldsymbol{z}_i^T \boldsymbol{b}_B)},$$

in this simulation study we took $\boldsymbol{z} = \boldsymbol{x}$. The general model formulation (Section 2.2) allows for $\boldsymbol{z}$ and $\boldsymbol{x}$ to be different. The value $\pi_{i,C} = 1 - \pi_{i,A} - \pi_{i,B}$.

In total three settings were explored which differ solely with regard to the values of $\boldsymbol{b}$ (which can be found in Table 2.1), resulting in different group sizes. In Setting I, there is a low number of "cured" (group C) cases of around 12%. Setting II contained around 35% cured cases, and Setting III around 67%. For all settings, the remainder of cases was approximately evenly split over groups A and B.

The event times differ for the two subgroups of A and B. Because of this, there are different parameter vectors for the survival functions of the subgroups. See Table 2.1 for the generating parameters $\boldsymbol{\beta}$. To model the subgroups, for each main group the observations are randomly split into two groups of equal size, with event times generated using a Weibull distribution each with a covariate vector $\boldsymbol{\beta}_{Ak}$ and corresponding scale parameter λ and shape parameter $1/\nu$, see Table 2.1. The event

					Setting I		Setting II		Setting III	
	$\boldsymbol{\beta}_{A1}$	$\boldsymbol{\beta}_{A2}$	$\boldsymbol{\beta}_{B1}$	$\boldsymbol{\beta}_{B2}$	$\boldsymbol{b}_A$	$\boldsymbol{b}_B$	$\boldsymbol{b}_A$	$\boldsymbol{b}_B$	$\boldsymbol{b}_A$	$\boldsymbol{b}_B$
(intercept)	-	-	-	-	0.7	0.2	-0.6	0.3	-0.7	-1.1
λ_j	0.7	0.9	0.5	0.8	-	-	-	-	-	-
ν_j	0.6	1	0.7	1.1	-	-	-	-	-	-
$x_{j,1}$	-0.2	0	0.3	1.3	0.7	-0.4	0.8	-1.2	0.5	-1.5
$x_{j,2}$	0.1	0.4	-0.5	-0.1	0.6	1.2	-0.1	-0.3	-1.2	-0.4
$x_{j,3}$	-0.5	-0.3	0.1	0.3	0.2	-1	0.6	-1	0.7	-1.3

Table 2.1: *Generating values for the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{b}$ in the simulation study.*

times for observations in the insusceptible group C are all the same, for these simulations we put these equal to 10^3. Finally, some censoring was introduced. For both simulation settings, the censoring distribution was Censor $\sim$ Unif$(0, 200)$, leading to $15 - 20\%$ censoring in the main groups A and B.

2.4.2 Simulation results

According to the simulation settings in Section 2.4.1, 1000 simulation runs were conducted. Three types of models are fitted and compared, (i) a homogeneous model with three main groups and no subgroups, K_j=1, (ii) a model with heterogeneity assuming two subgroups for both groups A and B, which corresponds to the data generation, and (iii) a model with too much heterogeneity as compared to how the data were generated, we use three subgroups for both main groups A and B.

Model without heterogeneity

We evaluate the estimation of the parameters in a misspecified model, assuming that there is no heterogeneity. Table 2.2 shows the mean of the parameter estimates over all simulation runs along with their standard

deviations. We observe that the parameters of the incidence model are well-estimated. The main interest, however, lies in the estimates $\hat{\boldsymbol{\beta}}$, since the model without heterogeneity forces to estimate only one $\boldsymbol{\beta}$ per main group, whereas the data are generated with two parameters $\boldsymbol{\beta}$ for both A and B. It is observed that abandoning the heterogeneity might lead to undetected effects. An example for this is $\boldsymbol{\beta}_{B,1}$: the model estimates are between 0.542 and 0.644 when using no subgroups, whereas the two generating values $\boldsymbol{\beta}_{B1,1} = 0.3$ and $\boldsymbol{\beta}_{B2,1} = 1.3$. It seems that the model favors weaker relationships, which results in estimated $\boldsymbol{\beta}$s that are relatively small in absolute value.

		$\hat{\boldsymbol{\beta}}_A$	$\hat{\boldsymbol{\beta}}_B$	$\hat{\boldsymbol{b}}_A$	$\hat{\boldsymbol{b}}_B$
Setting I	int.			0.862 (0.166)	0.386 (0.192)
	$x_{j,1}$	-0.083 (0.023)	0.542 (0.03)	0.643 (0.089)	-0.432 (0.101)
	$x_{j,2}$	0.13 (0.024)	-0.275 (0.022)	0.412 (0.07)	0.975 (0.087)
	$x_{j,3}$	-0.271 (0.018)	0.153 (0.017)	0.187 (0.082)	-0.977 (0.085)
Setting II	int.			-0.479 (0.125)	0.382 (0.15)
	$x_{j,1}$	-0.076 (0.03)	0.621 (0.039)	0.757 (0.05)	-1.16 (0.075)
	$x_{j,2}$	0.134 (0.023)	-0.262 (0.026)	-0.125 (0.031)	-0.353 (0.038)
	$x_{j,3}$	-0.257 (0.019)	0.163 (0.02)	0.549 (0.042)	-0.969 (0.046)
Setting III	int.			-0.591 (0.148)	-0.954 (0.182)
	$x_{j,1}$	-0.061 (0.042)	0.644 (0.053)	0.456 (0.049)	-1.426 (0.087)
	$x_{j,2}$	0.155 (0.038)	-0.257 (0.035)	-1.184 (0.05)	-0.454 (0.045)
	$x_{j,3}$	-0.238 (0.029)	0.168 (0.028)	0.628 (0.039)	-1.25 (0.053)

Table 2.2: *Mean and standard errors of the parameter estimates for the model without heterogeneity over 1000 simulation runs, for Setting I (with cure rate around* 12%*), Setting II (*35% *cure rate) and Setting III (*67% *cure rate).*

Model with heterogeneity: two subgroups

The mean parameter estimates for all settings over 1000 simulation runs for a model with heterogeneity can be found in Table 2.3. The param-

eter estimates for $\boldsymbol{b}$ in the right panel are close to the estimates for the model without heterogeneity, resulting in good estimates with respect to the simulated parameters and similar standard errors for all settings. The estimates $\hat{\boldsymbol{\beta}}$ in the left panel show that the two subgroups are well-identified for groups A and B. When comparing the estimates $\hat{\boldsymbol{\beta}}$ with the true parameter values used for simulation (see Table 2.1), we note that in general a higher cure rate does not disturb the estimation of $\boldsymbol{\beta}$. However, it should be noted that standard errors are larger for subgroup $A1$ when comparing with other subgroups. Additionally, standard errors tend to go up when the cured fraction is larger. This first phenomenon can be explained by the fact that the parameters of both subgroups in A lie closer to each other as compared to the subgroup parameters of B. This possibly makes estimation more difficult. The second phenomenon is explained by the smaller number of cases in groups A and B for Setting III in comparison with Setting I. Because approximately 2/3 of the observations is cured in Setting III, only around 600 observations are member of each subgroup in this setting. This leads to less accuracy and higher variation in parameter estimation.

In Table 2.4, the percentage of correctly classified observations for the model with two subgroups are listed on the diagonal of the middle part. When comparing these results with the percentages of correctly classified observations of the model without heterogeneity (on the left part of the same table), we observe a high percentage of correctly classified observations for all settings. Note that classification is better for groups A and B in the model with two subgroups, this is offset by a worse classification of the cured cases C, especially in Setting I. In general, we note that classification on the main group level does not incontestably favor one model over another. This is not an unexpected result, as classification is driven by the parameters $\boldsymbol{b}$, which only marginally change when changing the number of subgroups (see the $\boldsymbol{b}$-parameters in Tables 2.2 and 2.3). A simular result was observed when looking at the $\boldsymbol{b}$-parameters of a model when having three subgroups (see also Section 2.4.2).

The main advantage of the use of a heterogeneous model when there

		$\hat{\boldsymbol{\beta}}_{A1}$	$\hat{\boldsymbol{\beta}}_{A2}$	$\hat{\boldsymbol{\beta}}_{B1}$	$\hat{\boldsymbol{\beta}}_{B2}$	$\hat{\boldsymbol{b}}_{A}$	$\hat{\boldsymbol{b}}_{\boldsymbol{B}}$
Setting I	int.					0.877 (0.188)	0.385 (0.217)
	$x_{jk,1}$	-0.511 (0.378)	-0.056 (0.037)	0.36 (0.167)	1.05 (0.209)	0.635 (0.109)	-0.475 (0.123)
	$x_{jk,2}$	0.014 (0.296)	0.173 (0.078)	-0.39 (0.113)	-0.08 (0.135)	0.486 (0.091)	1.061 (0.106)
	$x_{jk,3}$	-0.997 (0.436)	-0.258 (0.029)	0.139 (0.073)	0.189 (0.106)	0.196 (0.107)	-0.991 (0.106)
Setting II	int.					-0.476 (0.138)	0.355 (0.16)
	$x_{jk,1}$	-0.51 (0.415)	-0.055 (0.09)	0.582 (0.198)	1.522 (0.325)	0.782 (0.064)	-1.214 (0.094)
	$x_{jk,2}$	0.087 (0.309)	0.151 (0.053)	-0.295 (0.067)	-0.132 (0.181)	-0.141 (0.038)	-0.353 (0.042)
	$x_{jk,3}$	-0.985 (0.364)	-0.235 (0.062)	0.158 (0.071)	0.311 (0.202)	0.598 (0.067)	-0.979 (0.057)
Setting III	int.					-0.57 (0.174)	-1.029 (0.201)
	$x_{jk,1}$	-0.323 (0.45)	-0.027 (0.129)	0.583 (0.125)	1.641 (0.406)	0.469 (0.055)	-1.496 (0.108)
	$x_{jk,2}$	0.284 (0.397)	0.208 (0.225)	-0.292 (0.07)	-0.106 (0.242)	-1.222 (0.064)	-0.445 (0.053)
	$x_{jk,3}$	-0.666 (0.385)	-0.208 (0.081)	0.148 (0.05)	0.393 (0.273)	0.67 (0.063)	-1.275 (0.064)

Table 2.3: *Mean and standard errors of the parameter estimates for the model with two subgroups for both A and B over 1000 simulation runs, for Setting I with low cure, Setting II with medium cure and Setting III with high cure.*

			Classified as (in %)								
			no heterogeneity			two subgroups			three subgroups		
			A	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*
Setting I	Real group	group *A*	94.74	0.56	4.70	95.55	0.65	3.79	95.63	0.66	3.71
		group *B*	0.66	94.70	4.64	0.80	95.80	3.40	0.82	95.82	3.36
		group *C*	12.89	9.50	77.62	17.95	13.95	68.09	18.60	14.06	67.34
Setting II	Real group	group *A*	92.64	0.15	7.21	93.58	0.15	6.26	93.66	0.16	6.18
		group *B*	0.16	91.03	8.80	0.21	91.66	8.14	0.21	91.74	8.04
		group *C*	4.87	5.61	89.52	7.43	6.85	85.72	7.72	7.03	85.25
Setting III	Real group	group *A*	89.08	0.05	10.87	90.07	0.06	9.87	95.63	0.66	3.71
		group *B*	0.05	88.65	11.30	0.06	89.45	10.48	0.82	95.82	3.36
		group *C*	1.36	2.21	96.43	2.38	2.78	94.84	18.60	14.06	67.34

Table 2.4: *Percentage of observations classified to each of the groups over 1000 simulation runs. The left part shows this for the model without heterogeneity, the middle part for the correct model with two subgroups for A and B, and the right part gives the classification percentages for the overspecified model with three subgroups for A and B. The percentages of correct classification are on the diagonals of each part.*

is some heterogeneity in the data lies in the possibility to model more accurate event times. To illustrate this, we investigate the estimates of the survival probabilities of the subjects using (i) a model without heterogeneity and (ii) with two subgroups per main group, and compare these estimates to the true survival probabilities of the simulated data. For the 1000 simulation runs, we considered the true survival rate in each subgroup (A1, A2, B1, B2), for each decile from 0.2 to 0.7 when sorting all 7500 event times. These are compared to the average estimated survival probability of all observations in groups A1, A2, B1 and B2, first using a homogeneous model, and secondly using a model with two subgroups. The result can be found in Table 2.5.

For each simulation setting at each of the six listed time deciles, the

		decile=2				decile=3				decile=4			
		$A1$	$A2$	$B1$	$B2$	$A1$	$A2$	$B1$	$B2$	$A1$	$A2$	$B1$	$B2$
Sett. I	1 group	**0.177**	0.164	**0.008**	*0.005*	**0.253**	**0.234**	**0.029**	*0.020*	**0.281**	**0.262**	**0.062**	*0.045*
	2 subgroups	**0.075**	*0.300*	**0.006**	*0.031*	**0.099**	**0.192**	**0.023**	*0.067*	**0.101**	**0.100**	**0.051**	*0.121*
Sett. II	1 group	**0.254**	*0.228*	*0.008*	*0.013*	**0.296**	**0.268**	**0.022**	*0.008*	**0.250**	**0.226**	**0.066**	*0.034*
	2 subgroups	**0.118**	*0.242*	*0.009*	*0.083*	**0.115**	**0.112**	**0.018**	*0.188*	**0.086**	**0.062**	**0.057**	0.335
Sett. III	1 group	**0.282**	**0.218**	**0.046**	*0.014*	**0.142**	0.054	**0.071**	*0.018*	0	0	0	0
	2 subgroups	**0.148**	**0.123**	**0.038**	*0.341*	**0.111**	*0.058*	**0.056**	*0.504*	0	0	0	0
		decile=5				decile=6				decile=7			
		$A1$	$A2$	$B1$	$B2$	$A1$	$A2$	$B1$	$B2$	$A1$	$A2$	$B1$	$B2$
Sett. I	1 group	**0.256**	**0.241**	**0.102**	*0.074*	**0.201**	**0.188**	**0.132**	*0.096*	**0.139**	**0.124**	**0.132**	*0.096*
	2 subgroups	**0.092**	**0.054**	**0.085**	*0.193*	**0.099**	**0.035**	**0.108**	*0.262*	**0.121**	**0.023**	**0.103**	*0.302*
Sett. II	1 group	**0.172**	**0.147**	**0.091**	*0.051*	*0.087*	**0.047**	**0.067**	*0.022*	0	0	0	0
	2 subgroups	**0.097**	**0.041**	**0.080**	*0.461*	*0.132*	**0.020**	**0.053**	*0.458*	0	0	0	0
Sett. III	1 group	0	0	0	0	0	0	0	0	0	0	0	0
	2 subgroups	0	0	0	0	0	0	0	0	0	0	0	0

Table 2.5: *The mean of the absolute differences between the population survival rate and estimated survival probabilities for the model without (1 group) and with (2 subgroups) heterogeneity for settings I–III. Six different time points are analyzed, looking at the deciles of the real event-times. Bold indicates a better performance for the heterogeneous model, italics indicates a better performance for the homogeneous model and regular print an equal performance. Zeroes are exact, both estimated survival probabilities and population survival rate are equal to zero.*

estimates of the proportion of the populations that experienced the event not later than time t_d, $S_{jk}(t_0) - S_{jk}(t_d) = 1 - S_{jk}(t_d)$, with t_d equal to the generated time decile d and t_0 the starting time of the study (hence, before any event has occurred) are compared to the true proportions. Table 2.5 lists the absolute differences between the estimated and the true proportions, either using a model with or without heterogeneity. When the model without heterogeneity performs better than the model with heterogeneity (this is, when the absolute difference for "1 group" is lower than for "2 subgroups"), the numbers in Table 2.5 are indicated in italics, the numbers in bold represent situations where the model with heterogeneity performed better. Note that, for high censoring rates and for bigger time deciles, the survival probability estimate goes to 0 for both models, and both models perform about equally (regular text font). Note that in 75% of the cases, the model with heterogeneity performed either equally well or better than the model without heterogeneity.

Model with heterogeneity: three subgroups

The third model that was fit to the simulated data was a model with three subgroups for both main groups A and B. For this model, it appears that in many of the simulation runs, the $\hat{\boldsymbol{\beta}}_{jk}$ for several subgroups are converging to (approximately) equal values, see Table 2.6. This is a result of putting several subgroup probabilities equal to $1/K_j$, as discussed in Section 2.3.2. In 606, 737 and 756 out of the 1000 simulation runs (for Setting I to III respectively), at least two out of three $\boldsymbol{\beta}_{jk}$ for groups A, B or both were estimated to be equal, with equal shares of observations classified in those equally estimated groups. The occurrence of equal parameter estimates is a strong indication that too many subgroups were modeled and that the number of subgroups K_j should be decreased. In the data example, we use this in combination with a version of Akaike's information criterion to determine the number of subgroups.

group A	group B	Setting I	Setting II	Setting III
$\#(\hat{\boldsymbol{\beta}}_{A1}, \hat{\boldsymbol{\beta}}_{A2}, \hat{\boldsymbol{\beta}}_{A3}) = 3$	$\#(\hat{\boldsymbol{\beta}}_{B1}, \hat{\boldsymbol{\beta}}_{B2}, \hat{\boldsymbol{\beta}}_{B3}) = 3$	394	263	244
$\#(\hat{\boldsymbol{\beta}}_{A1}, \hat{\boldsymbol{\beta}}_{A2}, \hat{\boldsymbol{\beta}}_{A3}) = 2$	$\#(\hat{\boldsymbol{\beta}}_{B1}, \hat{\boldsymbol{\beta}}_{B2}, \hat{\boldsymbol{\beta}}_{B3}) = 3$	171	125	164
$\#(\hat{\boldsymbol{\beta}}_{A1}, \hat{\boldsymbol{\beta}}_{A2}, \hat{\boldsymbol{\beta}}_{A3}) = 3$	$\#(\hat{\boldsymbol{\beta}}_{B1}, \hat{\boldsymbol{\beta}}_{B2}, \hat{\boldsymbol{\beta}}_{B3}) = 2$	315	438	363
$\#(\hat{\boldsymbol{\beta}}_{A1}, \hat{\boldsymbol{\beta}}_{A2}, \hat{\boldsymbol{\beta}}_{A3}) = 2$	$\#(\hat{\boldsymbol{\beta}}_{B1}, \hat{\boldsymbol{\beta}}_{B2}, \hat{\boldsymbol{\beta}}_{B3}) = 2$	120	174	229

Table 2.6: *Analysis of the parameter estimates for heterogeneity with 3 subgroups for groups A and B, for censoring settings I–III.* $\#(\ldots) = 3$ *denotes that all three parameter estimates are different,* $\#(\ldots) = 2$ *denotes that two out of three parameter estimates are the same. In the majority of the simulation runs, there were equal estimates for two out of the three* $\hat{\boldsymbol{\beta}}_{jk}$ *for A or B, or both.*

2.5 Data example on credit risk

2.5.1 Data description

We analyze a credit loan data set, with the main interest in the prediction of defaults and early loan repayments. The cured or insusceptible group is given by the matured loans with the loan repayment on the predefined end date (maturity). The data are from a UK bank, previously used in Stepanova and Thomas (2002a) and Dirick et al. (2015). This data set consists of 7521 observations with loan term 36 months and 13 variables. Table 2.7 lists the eight variables that were used to build our model. The event of early repayment was observed 2992 times, default 376 times and maturity 269 times. The remaining 3884 other observations were censored.

2.5.2 Decision on the number of subgroups

To determine the value of K_j for each of the $J - 1{=}2$ main groups, or rephrased to this data set, the K_d subgroups for default and the K_p subgroups for early repayment, we use a version of Akaike's information criterion (AIC) that accounts for incomplete data. Introduced by Cavanaugh and Shumway (1998), the so-called "complete-data AIC" (AIC_{cd}) makes use of the expected complete-data log likelihood instead of the observed

	Description	Type	**default**	**early repayment**	**cured**	**unknown (censored)**
x_1	The gender of the customer (1=M, 0=F)	cat	75.0%	71.2%	71.7%	73.6%
x_2	Amount of the loan	cont	3588.8 (1842.4)	3607.4 (1756.8)	3740.9 (1900.2)	3630.7 (1805.9)
x_3	Number of years at current address	cont	6.0 (6.9)	7.2 (7.5)	8.2 (8.0)	8.2 (7.9)
x_4	Number of years at current employer	cont	4.5 (4.8)	7.6 (7.4)	9.2 (9.0)	8.1 (7.7)
x_5	Amount of insurance premium	cont	342.3 (293.0)	219.3 (260.0)	200.9 (260.0)	231.8 (272.8)
x_6	Home phone or not (1=N,0=Y)	cat	6.9%	3.7%	3.7%	4.2%
x_7	Own house or not (1=N, 0=Y)	cat	46.5%	31.4%	26.8%	32.9%
x_8	Payment frequency (1=low/unknown, 0=high)	cat	56.6%	69.5%	66.9%	67.0%

Table 2.7: *Data on credit risk. Description of the variables, continuous (cont) or categorical (cat), stratified by failure event. For continuous variables, the observed mean (and standard deviation) is given, for categorical variables (which are all binary) the proportion of one-values.*

log likelihood. Dirick et al. (2015) obtained the AIC_{cd} for multiple event mixture cure models. In this context,

$$AIC_{cd} = -2Q^h\big(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}\big) + 2d + 2 \text{ trace}\{\boldsymbol{DM}(I_d - \boldsymbol{DM})^{-1}\}.$$

An additional selection restriction is that increasing the number of subgroups K_j should be stopped when the estimates $\hat{\boldsymbol{\beta}}_{jk}$ of different components are the same.

For the data on credit risk, we considered values for both K_p and K_d in $\{1, 2, 3\}$, with the value of one representing homogeneity. Using all combinations for K_p and K_d gave rise to 9 models. Figure 2.1 graphically represents the AIC_{cd}-values, with a minimum AIC_{cd} for a model where there is no heterogeneity for default, but 3 heterogeneity-groups for early repayment. When looking at the estimated values of the $\boldsymbol{\beta}_{pk}$, however, we received equal estimates for two out of three parameter vectors. This was also the case for $\boldsymbol{\beta}_{pk}$ estimates for ($K_p = 3$, $K_d = 2$) and ($K_p = 3$, $K_d = 3$), which have the next lowest AIC_{cd}-values. The next lowest AIC_{cd}-value has K_p=2, $K_d = 1$, which is the preferred value as there we have the model with minimal AIC_{cd} without equal estimates for $\boldsymbol{\beta}_{pk}$-parameters. Hence, the suggested final model is one with only one group for default, and two subgroups for early repayment.

Note that the need for an adequate information criterion becomes apparent, as the AIC_{cd} is an appropriate criterion in presence of incomplete data, but is not able to identify the presence of two (or more) parameter vectors that are equal in a mixture setting. One could add an extra penalty when equal estimates occur, but doing this, the criterion loses its $AICcd$-properties. Naik et al. (2007) derive a so-called "Mixture Regression Criterion" to perform model selection in precense of several mixture components. An extension of that criterion that would allow for censored survival data in the mixture cure models could be a possibility.

2.5.3 Final result

The parameter estimates of the final model are given in Table 2.8, along with their standard errors and p-values. Using these results, a reduced

	$\hat{\boldsymbol{\beta}}_{d1}$	$\hat{\boldsymbol{\beta}}_{p1}$	$\hat{\boldsymbol{\beta}}_{p2}$	$\hat{\boldsymbol{b}}_d$	$\hat{\boldsymbol{b}}_p$
τ	1	0.587	0.413		
int.				-1.29^{***} (0.211)	0.535^{***} (0.113)
x_1	0.245^{*} (0.123)	0.0223 (0.054)	0.156^{*} (0.065)	-0.120 (0.139)	-0.190^{**} (0.073)
x_2	$-7 \times 10^{-5*}$ (3×10^{-5})	2×10^{-5} (1×10^{-5})	$-9 \times 10^{-5***}$ (2×10^{-5})	-3×10^{-5} (4×10^{-5})	2×10^{-5} (2×10^{-5})
x_3	-0.022^{**} (0.008)	-0.013^{***} (0.003)	-0.009^{*} (0.004)	-0.030^{***} (0.009)	-0.0128^{***} (0.004)
x_4	-0.051^{***}(0.012)	-2.14e-4 (0.003)	-0.004 (0.004)	-0.067^{***} (0.012)	-0.011^{*} (0.004)
x_5	4×10^{-4} (2×10^{-4})	$-3 \times 10^{-4**}$ (9×10^{-5})	-1×10^{-4} (1×10^{-4})	0.001^{***} (2×10^{-4})	1×10^{-4} (1×10^{-4})
x_6	0.56^{**} (0.206)	-0.234 (0.125)	-0.42^{**} (0.157)	0.393 (0.247)	0.123 (0.179)
x_7	-0.0423 (0.108)	-0.0201 (0.054)	-0.192^{**} (0.0652)	0.408^{***} (0.123)	0.106 (0.073)
x_8	-0.032 (0.11)	0.090 (0.054)	0.22^{**} (0.065)	-0.38^{**} (0.126)	-0.066 (0.072)

Table 2.8: *Data on credit risk. Parameter estimates (standard errors) for the hierarchical mixture cure model with $K_d = 1$, $K_p = 2$. The value τ represents the proportion of the population belonging to a respective subgroup, given the main group, 'int.' stands for intercept. Because of asymptotic normality, the standard errors are used to obtain p-values. * denotes significance at the 0.05-level, ** significance at 0.01-level, and *** significance at the 0.001-level.*

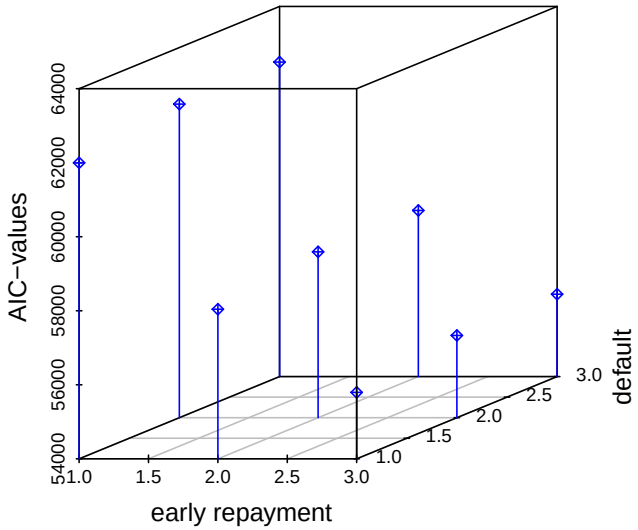

Figure 2.1: *Data on credit risk. The AIC_{cd}-values of the hierarchical models for the number of subgroups for early repayment and default varying between 1 and 3.*

model including only significant covariates could be constructed, leading to different vector lengths for $\hat{\boldsymbol{\beta}}_{d1}, \hat{\boldsymbol{\beta}}_{p1}$ and $\hat{\boldsymbol{\beta}}_{p2}$. Note that vector lengths of $\hat{\boldsymbol{b}}_d$ and $\hat{\boldsymbol{b}}_p$ are theoretically constrained to be of equal length, as they are jointly estimated in a multinomial logit. Setting III of the simulation study is similar with regard to the approximately 50% censoring present in this dataset. Hence we seem to get consistent estimates for $\boldsymbol{\beta}$, but for $\boldsymbol{b}$, there might be small deviations due to the relatively high censoring percentage. For the latter parameter group, mainly the sign and the relative magnitude should be used the analysis.

First, we discuss $\hat{\boldsymbol{b}}_d$ and $\hat{\boldsymbol{b}}_p$ which effect the probabilities of default and prepayment, which are smaller for men than for women. The effect on prepayment is stronger and statistically significant. The residential stability (x_3) and employment stability (x_4) reduce significantly both probabilities,

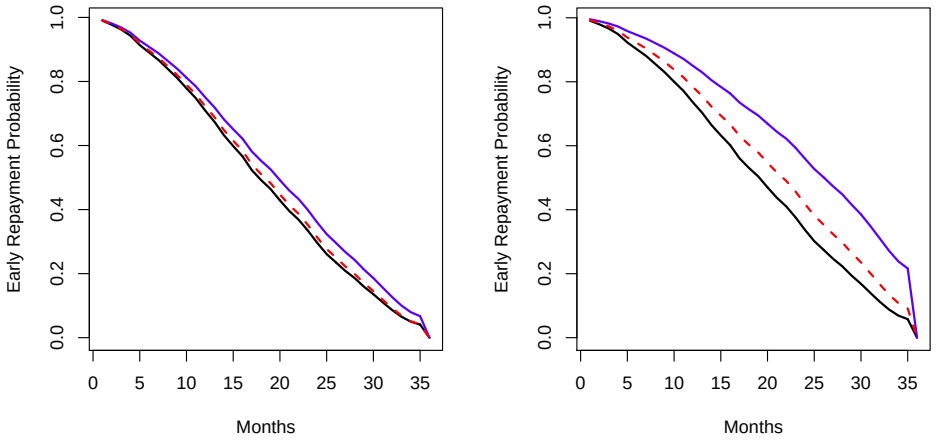

Figure 2.2: *Credit loan data. Estimated survival curves for two random observations (one in the left and another one in the right panel). The black and blue lines are, respectively, the survival curves (estimated through formula (2.3) and the Breslow-type estimator for the baseline hazard) for early repayment group 1 and group 2, for the final model where heterogeneity is present in the early repayment-group. The red dotted line represents the estimated survival curve fitted with a model with no heterogeneity (hence* $\widehat{S}(t \mid \tilde{Y} = p; \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_p)$*, assuming no subgroups).*

but the effect on the probability of default is much stronger. Having no home phone (x_6) and no own home (x_7) greatly increase both probabilities, the relative effect on the default probabilities is even stronger here. However only the coefficient of x_7 for default is statistically significant. The low frequency of payment reduces significantly the default probability. To summarize, the variables affect the probability of defaults more than the probability of prepayment and the signs of the coefficients are reasonable.

Second, we discuss the parameters of the survival function, $\hat{\boldsymbol{\beta}}_{d1}$, $\hat{\boldsymbol{\beta}}_{p1}$ and $\hat{\boldsymbol{\beta}}_{p2}$. When men go into default and prepayment the time to event is much shorter, though it is statistically significant only for the second pre-

payment group. The residential stability (x_3) and employment stability (x_4) delay the timing of defaults and prepayments and mostly statistically significant. Having no home phone (x_6) accelerates defaults, but delays the prepayments. Having no own home (x_6) significantly delays the prepayment of the second default group. The low frequency of payments delays the defaults, but accelerates the prepayments. The variables can have different effects in the probability of the event and the conditional survival function. The gender dummy variable (x_1) decreases the probability of prepayment or default for men, but if men experience the event, it happens faster. No home phone (x_6) and no own home (x_7) increases the prepayment probability, but delays the timing of prepayment. Finally, when two prepayment groups are compared, all coefficients for the second group are larger. The first group can be called the 'base group', while the second group can be called the 'sensitive' group. All factors have much stronger effects in the 'sensitive' group. To illustrate the difference in terms of early repayment probabilities between the base group and the sensitive group, the two curves representing early repayment for two of the observations are plotted in Figure 2.2 (for one random observation in the left panel and for the other one in the right panel). The black and blue curves represent early repayment groups 1 and 2 respectively. The red dotted curve represents the early repayment curve for the two observations fitted using a model without heterogeneity. This figure illustrates that the results for unobserved heterogeneity can differ quite a bit from one observation to another: where for the observation represented in the left panel the black and blue curve lie relatively close, the differences are substantial for the observation represented in the right panel. Note that the two curves can be further apart for some observations (as on the right panel), and closer for other observations (as on the left panel).

2.6 Conclusion

This chapter highlights the importance of the unobserved heterogeneity for credit risk models. It derives a hierarchical EM algorithm for estimation

and provides the first simulation study in this area, which reveals that not using models that allow heterogeneity when heterogeneity is present can lead to distorted conclusions on the magnitude of the parameters related to the timing of a certain event.

An application of the model for the loan data from a UK bank finds that the explanatory variables can act in different directions upon incidence and duration; and, variables exist that are statistically significant in explaining only incidence or duration.

While the model proposed is general and can be used in many contexts, there are still many aspects that would be interesting for future research. First of all, it would be interesting to consider other survival analysis techniques for cure models for extension towards heterogeneity, such as the promotion time cure model, or incorporating nonparametric effects. The presence of censoring and its effects on estimation also requires more investigation, one possibility is to explicitly incorporate censoring as missing information in the EM-algorithm.

2.7 Appendix to Chapter 2

2.7.1 Identifiability of the main groups of the hierarchical mixture cure model.

Similar as in Heckman and Honoré (1989); Peng and Zhang (2008), we start from (2.2). For a model with three main groups, without subgroups, the general survival function, unconditioal on Y_j but for given values of $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{b}$ and $\boldsymbol{\beta}$, can be written as

$$S(t;\boldsymbol{x},\boldsymbol{z},\boldsymbol{b},\boldsymbol{\beta}) = \sum_{j=1}^{2} \pi_j(\boldsymbol{z},\boldsymbol{b}_j) S(t\,|\,Y_j=1;\boldsymbol{x},\boldsymbol{\beta}_j) + 1 - \sum_{j=1}^{2} \pi_j(\boldsymbol{z},\boldsymbol{b}_j) \quad (2.10)$$

where

$$\begin{aligned} S(t \mid Y_j=1;\boldsymbol{x},\boldsymbol{\beta}_j) &= \exp\Big\{ -\exp(\boldsymbol{x}^T\boldsymbol{\beta}_j) \int_0^t h_0(u \mid \widetilde{Y}_j=1)du \Big\} \\ &\equiv \exp\Big(-\phi_j(\boldsymbol{x}) H_0^j(t)\Big), \end{aligned}$$

with $\exp(\boldsymbol{x}^T\beta_j) = \phi_j(\boldsymbol{x})$ for $j = 1, 2$. Omitting parameters for notational convenience,

$$S(t; \boldsymbol{x}, \boldsymbol{z}) = 1 + \sum_{j=1}^{2} \{\pi_j(\boldsymbol{z}) \exp\{-\phi_j(\boldsymbol{x}) H_0^j(t)\} - \pi_j(\boldsymbol{z})\}. \tag{2.11}$$

Theorem 1. *Assume that for* $j = 1, 2$,
(A1) The cumulative hazard function satisfies $\lim_{t\to 0} H_0^j(t) = H_0^j(0) = 0$,
(A2) $\pi_j(\boldsymbol{z})$ *is non-negative and non-constant*
(A3) $\phi_j(\boldsymbol{x})$ *is non-negative, differentiable and non-constant with* $\phi_j(0) = 1$.
Then, $S(t; \boldsymbol{x}, \boldsymbol{z})$ *as in (2.10) is identifiable.*
This means that if $\{\pi_j(\boldsymbol{z}), \phi_j(\boldsymbol{x}), H_0^j; j = 1, 2\}$, *and* $\{\pi_j^*(\boldsymbol{z}), \phi_j^*(\boldsymbol{x}), H_0^{j*}; j = 1, 2\}$ *are such that*

$$\begin{aligned} &\sum_{j=1}^{2} \{\pi_j(\boldsymbol{z}) \exp\Big(-\phi_j(\boldsymbol{x}) H_0^j(t)\Big) - \pi_j(\boldsymbol{z})\} \\ &\quad = \sum_{j=1}^{2} \{\pi_j^*(\boldsymbol{z}) \exp\Big(-\phi_j^*(\boldsymbol{x}) H_0^{*j}(t)\Big) - \pi_j^*(\boldsymbol{z})\} \end{aligned} \tag{2.12}$$

then it follows that $\pi_j(\boldsymbol{z}) = \pi_j^*(\boldsymbol{z})$, $\phi_j(\boldsymbol{x}) = \phi_j^*(\boldsymbol{x})$, $H_0^j = H_0^{*j}$, *for* $j = 1, 2$.

Proof. Denote T_1 the timing of event type 1, T_2 the timing of event type 2, and only the event type that happens first is actually observed. We define 'crude' survival functions $K_1(t)$ and $K_2(t)$ as the probability of not experiencing event type 1, resp. 2, before time t,

$$\begin{aligned} K_1(t; \boldsymbol{x}, \boldsymbol{z}) &= P((T_1 > t) \cap (T_2 > T_1); \boldsymbol{x}, \boldsymbol{z}), \\ K_2(t; \boldsymbol{x}, \boldsymbol{z}) &= P((T_2 > t) \cap (T_1 > T_2); \boldsymbol{x}, \boldsymbol{z}). \end{aligned}$$

Similar to Tsiatis (1975), and from (2.10), we obtain that for $j = 1, 2$,

$$\frac{\partial K_j}{\partial t}(t \mid \boldsymbol{x}, \boldsymbol{z}) = \frac{\partial}{\partial t}\Big(\pi_j(\boldsymbol{z}, \boldsymbol{b}_j) S(t \mid Y_j = 1; \boldsymbol{x}, \boldsymbol{\beta}_j)\Big).$$

Fix $j \in \{1, 2\}$. From (2.11), it follows that

$$\frac{\partial K_j}{\partial t}(t) = -\pi_j(\boldsymbol{z}) \phi_j(\boldsymbol{x}) h_0^j(t) \exp\Big(-\phi_j(\boldsymbol{x}) H_0^j(t)\Big) \equiv s^j(t; \boldsymbol{x}, \boldsymbol{z}). \tag{2.13}$$

First we show that there exists a constant c_j such that

$$h_0^{*j}(t)\exp\left(-\phi_j^*(\boldsymbol{x})H_0^{*j}(t)\right) = c_j h_0^j(t)\exp\left(-\phi_j(\boldsymbol{x})H_0^j(t)\right). \tag{2.14}$$

Case 1: $\boldsymbol{x} = \boldsymbol{z}$. Take any constant x_0 in the domain of $s^1(\cdot;\boldsymbol{x})$. Dividing $s^j(t;\boldsymbol{x})$ by $s^j(t;\boldsymbol{x_0})$, we get

$$\begin{aligned}\frac{s^j(t;\boldsymbol{x})}{s^j(t;\boldsymbol{x_0})} &= \frac{-\pi_j(\boldsymbol{x})\phi_j(\boldsymbol{x})h_0^j(t)\exp\left(-\phi_j(\boldsymbol{x})H_0^j(t)\right)}{-\pi_j(\boldsymbol{x_0})\phi_j(\boldsymbol{x_0})h_0^j(t)\exp\left(-\phi_j(\boldsymbol{x_0})H_0^j(t)\right)} \\ &= \frac{-\pi_j(\boldsymbol{x})\phi_j(\boldsymbol{x})\exp\left(-\phi_j(\boldsymbol{x})H_0^j(t)\right)}{-\pi_j(\boldsymbol{x_0})\phi_j(\boldsymbol{x_0})\exp\left(-\phi_j(\boldsymbol{x_0})H_0^j(t)\right)}\end{aligned}$$

and letting $t \to 0$ we get by assumption (A1) ($\lim_{t\to 0} H_0^j(t) = H_0^j(0) = 0$)

$$-\pi_j(\boldsymbol{x_0})\phi_j(\boldsymbol{x_0}) = -\pi_j(\boldsymbol{x})\phi_j(\boldsymbol{x})\lim_{t\to 0}\frac{s^j(t;\boldsymbol{x})}{s^j(t;\boldsymbol{x_0})}.$$

Likewise, $-\pi_j^*(\boldsymbol{x_0})\phi_j^*(\boldsymbol{x_0}) = -\pi_j^*(\boldsymbol{x})\phi_j^*(\boldsymbol{x})\lim_{t\to 0}\frac{s^j(t;\boldsymbol{x})}{s^j(t;\boldsymbol{x_0})}$. Consequently,

$$\frac{\pi_j(\boldsymbol{x_0})\phi_j(\boldsymbol{x_0})}{\pi_j^*(\boldsymbol{x_0})\phi_j^*(\boldsymbol{x_0})} = \frac{\pi_j(\boldsymbol{x})\phi_j(\boldsymbol{x})}{\pi_j^*(\boldsymbol{x})\phi_j^*(\boldsymbol{x})} \equiv c_j; \tag{2.15}$$

hence $\pi_j(\boldsymbol{x})\phi_j(\boldsymbol{x})$ can be determined upon a constant. Differentiating (2.13) with respect to t then leads to

$$\begin{aligned}&-\pi_j(\boldsymbol{x})\phi_j(\boldsymbol{x})h_0^j(t)\exp\left(-\phi_j(\boldsymbol{x})H_0^j(t)\right) \\ &\quad = -\pi_j^*(\boldsymbol{x})\phi_j^*(\boldsymbol{x})h_0^{*j}(t)\exp\left(-\phi_j^*(\boldsymbol{x})H_0^{*j}(t)\right)\end{aligned}$$

and, using (2.15)

$$h_0^{*j}(t)\exp\left(-\phi_j^*(\boldsymbol{x})H_0^{*j}(t)\right) = c_j h_0^j(t)\exp\left(-\phi_j(\boldsymbol{x})H_0^j(t)\right).$$

Case 2: $\boldsymbol{x} \neq \boldsymbol{z}$. By dividing $s^j(t;\boldsymbol{x},\boldsymbol{z})$ by $s^j(t;0,\boldsymbol{z})$ and letting $t \to 0$ gives by assumptions A1, and A3, that $\lim_{t\to 0}\frac{s^j(t;\boldsymbol{x},\boldsymbol{z})}{s^j(t;0,\boldsymbol{z})} = \phi_j(\boldsymbol{x})$. Hence $\phi_j(\boldsymbol{x})$ is

uniquely determined. Consider such another set $\{\pi_j^*(\boldsymbol{z}), \phi_j(\boldsymbol{x}), H_0^{j*};$ $j = 1, 2\}$. Then, for any value $\boldsymbol{z}_0$ in the domain of s^j, dividing $s^j(t; \boldsymbol{x}, \boldsymbol{z})$ by $s^j(t; \boldsymbol{x}, \boldsymbol{z_0})$ and letting $t \to 0$ leads to $\lim_{t\to 0} \dfrac{s^j(t; \boldsymbol{x}, \boldsymbol{z})}{s^j(t; \boldsymbol{x}, \boldsymbol{z_0})} = \dfrac{\pi_j(\boldsymbol{z})}{\pi_j(\boldsymbol{z_0})} = \dfrac{\pi_j^*(\boldsymbol{z})}{\pi_j^*(\boldsymbol{z_0})}$. Hence, $\dfrac{\pi_j^*(\boldsymbol{z})}{\pi_j(\boldsymbol{z})}$ must be a constant, which we define as c_j. Using these results, similarly to the case $x = z$, Differentiating (2.13) with respect to t then leads to

$$\begin{aligned} &-\pi_j(\boldsymbol{z})\phi_j(\boldsymbol{x})h_0^j(t)\exp\left(-\phi_j(\boldsymbol{x})H_0^j(t)\right) \\ &\quad = -\pi_j^*(\boldsymbol{z})\phi_j^*(\boldsymbol{x})h_0^{*j}(t)\exp\left(-\phi_j^*(\boldsymbol{x})H_0^{*j}(t)\right) \end{aligned}$$

and

$$h_0^{*j}(t)\exp\left(-\phi_j^*(\boldsymbol{x})H_0^{*j}(t)\right) = c_j h_0^j(t)\exp\left(-\phi_j(\boldsymbol{x})H_0^j(t)\right).$$

Let $t \to 0$ on both sides of (2.14), then we get $\lim_{t\to 0} \dfrac{h_0^{*j}(t)}{h_0^j(t)} = c_j$, which is well-defined. When we take the derivatives on both sides of (2.14) with respect to $\boldsymbol{x}$, we obtain by (A3),

$$\begin{aligned} &h_0^{*j}(t)H_0^{*j}(t)\frac{\partial\phi_j^*(\boldsymbol{x})}{\partial\boldsymbol{x}}\exp\Big(-\phi_j^*(\boldsymbol{x})H_0^{*j}(t)\Big) \\ &\quad = c_j h_0^j(t)H_0^j(t)\frac{\partial\phi_j(\boldsymbol{x})}{\partial\boldsymbol{x}}\exp\Big(-\phi_j(\boldsymbol{x})H_0^j(t)\Big). \end{aligned} \tag{2.16}$$

Let $t \to 0$ on both sides of (2.16). Since $\lim_{t\to 0} \dfrac{H_0^j(t)}{H_0^{*j}(t)} = \dfrac{1}{c_j} = \lim_{t\to 0} \dfrac{h_0^j(t}{h_0^{*j}(t)}$,

$$\frac{\partial\phi_j^*(\boldsymbol{x})}{\partial\boldsymbol{x}} \Big/ \frac{\partial\phi_j(\boldsymbol{x})}{\partial\boldsymbol{x}} = \frac{1}{c_j}. \tag{2.17}$$

Integrating equation (2.17) with respect to x, we get by (A3)

$$\phi_j^*(\boldsymbol{x}) = \frac{1}{c_j}\phi_j(\boldsymbol{x}) - \frac{1}{c_j} + 1. \tag{2.18}$$

Take $x = 0$. Because $\phi(0) = \phi^*(0) = 1$ from (A3), (2.14) simplifies to

$$c_j h_0^j(t)\exp\left(-H_0^j(t)\right) = h_0^{*j}(t)\exp\left(-H_0^{*j}(t)\right). \tag{2.19}$$

From the ratios of (2.14) and (2.19), we get that $H_0^{*1}(\boldsymbol{x})(\phi_1^*(\boldsymbol{x})-1) = H_0^1(\boldsymbol{x})(\phi_1(\boldsymbol{x})-1)$. Using (2.18), it is then easy to show that $H_0^{*j}(t) = c_j H_0^j(t)$ and consequently $h_0^{*j}(t) = c_j h_0^j(t)$. From (2.19) follows that $H_0^{*1}(t) = H_0^1(t)$ and $c_1 = 1$. In addition, we obtain $\phi_j(\boldsymbol{x}) = \phi_j^*(\boldsymbol{x})$ and $\pi_j(\boldsymbol{z}) = \pi_j^*(\boldsymbol{z})$. □

2.7.2 Relationship between $\log \tau_{\tilde{y}|j}$ and $v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)$

From expression (2.6), the second term to be maximized is

$$\sum_{i=1}^{n} E[\log \tau_{\widetilde{Y}_i|Y_i} \,|\, T_i; \boldsymbol{\beta}^{(r)}_{Y_i,\widetilde{Y}_i}] = \sum_{i=1}^{n}\sum_{y=1}^{J}\sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|y}(\boldsymbol{\beta}_y^{(r)}; t_i, \boldsymbol{x}_i) w_y(\boldsymbol{\beta}^{(r)}; t_i, \boldsymbol{x}_i) \log \tau_{\tilde{y}|y}.$$

Conditioning on the main group, say j, the term with $y = j$ in the above sum equals

$$\sum_{i=1}^{n}\sum_{\tilde{y}=1}^{K_j-1} v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i) \log \tau_{\tilde{y}|j} + \sum_{i=1}^{n} v_{K_j|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i) \log(1 - \sum_{\tilde{y}=1}^{K_j-1} \tau_{\tilde{y}|j}).$$

Setting the partial derivative with respect to $\tau_{\tilde{y}|j}$ equal to 0,

$$\begin{aligned} &\frac{\partial Q^h\big((\boldsymbol{b}, \boldsymbol{\beta}_{Y,\widetilde{Y}})^{(r+1)} \mid (\boldsymbol{b}, \boldsymbol{\beta}_{Y,\widetilde{Y}})^{(r)}\big)}{\partial \tau_{\tilde{y}|j}} \\ &= \sum_{i=1}^{n} \frac{v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)}{\tau_{\tilde{y}|j}} - \sum_{i=1}^{n} \frac{v_{K_j|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)}{\tau_{K_j|j}} = 0, \end{aligned}$$

implies that the optimizer $\tau_{\tilde{y}|j}^{(r+1)}$ satisfies

$$\tau_{\tilde{y}|j}^{(r+1)} = \frac{\sum_{i=1}^{n} v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)}{\sum_{i=1}^{n} v_{K_j|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)} \tau_{K_j|j}^{(r+1)} \tag{2.20}$$

for $\tilde{y} = 1, \ldots, K_j - 1$. Under the constraints that for every $j = 1 \ldots, K$ the weights $v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i); \tilde{y} = 1 \ldots, K_j$ sum to 1, we obtain

$$1 = \sum_{\tilde{y}=1}^{K_j} \tau_{\tilde{y}|j}^{(r+1)} = \frac{\sum_{\tilde{y}=1}^{K_j}\sum_{i=1}^{n} v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i) \tau_{K_j|j}^{(r+1)}}{\sum_{i=1}^{n} v_{K_j|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^{n}\left(\sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)\right)\tau_{K_j|j}^{(r+1)}}{\sum_{i=1}^{n} v_{K_j|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)} \\
&= \frac{\sum_{i=1}^{n}\tau_{K_j|j}^{(r+1)}}{\sum_{i=1}^{n} v_{K_j|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)} = \frac{n\tau_{K_j|j}^{(r+1)}}{\sum_{i=1}^{n} v_{K_j|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)}.
\end{aligned}
$$

So, $\tau_{K_j|j}^{(r+1)} = n^{-1}\sum_{i=1}^{n} v_{K_j|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)$. When plugging this back in (2.20), the same form follows for $\tilde{y} = 1, \ldots, K_j - 1$. Because of this, $\tau_{\tilde{y}|j}^{(r+1)} = n^{-1}\sum_{i=1}^{n} v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)$. With the constraint that $\sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|y}(\boldsymbol{\beta}_y^{(r)}; t_i, \boldsymbol{x}_i) = 1$ for each i, these $v_{\tilde{y}|j}(\boldsymbol{\beta}_j^{(r)}; t_i, \boldsymbol{x}_i)$ are well defined, hence so is $\tau_{\tilde{y}|j}^{(r+1)}$.

Chapter 3

Time to default in credit scoring using survival analysis: a benchmark study

Abstract

We investigate the performance of various survival analysis techniques applied to ten actual credit data sets from Belgian and UK financial institutions. In the comparison we consider classical survival analysis techniques, namely the accelerated failure time models and Cox proportional hazards regression models, as well as Cox proportional hazard regression models with splines in the hazard function. Mixture cure models for single and multiple events were more recently introduced in the credit risk context. The performance of these models is evaluated using both a statistical evaluation and an economic approach through the use of annuity theory. It is found that spline-based methods and the single event mixture cure model perform well in the credit risk context.

This chapter is based on Dirick, L., Claeskens, G. and Baesens, B. (2015). Time to default in credit scoring using survival analysis: a benchmark study. Working paper, submitted.

3.1 Introduction

With the introduction of compliance guidelines such as Basel II and Basel III, and the resulting higher need for more accurate credit risk calculations, survival analysis gained more importance over the recent years. Historically, survival analysis is mainly used in the medical context as well as in engineering, where the time duration until an event is analyzed, for example the time until death or machine failure (See Kalbfleisch and Prentice, 2002; Collett, 2003; Cox and Oakes, 1984).

As an alternative to logistic regression, Narain (1992) first introduced the idea of using survival analysis in the credit risk context. The advantage of using survival analysis in this context is that the time to default can be modeled, and not just whether an applicant will default or not (Thomas et al., 2002). Many authors followed the example of Narain (1992) and started to use more advanced methods as compared to the parametric accelerated failure time (AFT) survival methods used in this first work. An overview is given in Table 3.1. With its flexible non-parametric baseline hazard, the Cox proportional hazard (Cox PH) model was an obvious first alternative to the AFT model (Banasik et al., 1999), and subsequent contributions extended both Cox PH and AFT models by using, among others, coarse classification (Stepanova and Thomas, 2002b) and time-varying covariates (Bellotti and Crook, 2009). In recent research, mixture cure models have become popular in this context, as these models allow to model a "cured" fraction, a part of the population that will not go into default, in survival models.

In the existing literature, some questions remain. Firstly, except for Zhang and Thomas (2012), there has been no an attempt to compare a wide range of the available methods in one paper. Secondly, in each of the papers listed in Table 3.1, only one data set was analyzed, not allowing to draw conclusions on which of the survival methods to use. Finally, in the majority of the papers, the evaluation remains largely focused on classification and the area under the curve (AUC) of the receiver operating characteristics curve. In this chapter, we contribute to the existing litera-

ture by analyzing ten different data sets from five banks, using all classes of models listed in Table 3.1, and using both statistical (AUC and predicted time error measurement) and economic evaluation measures (by predicting the future value of the loan), applicable to all model types considered, the "plain" survival models as well as the mixture cure models.

The remainder of this chapter is organized as follows. Section 3.2 gives an overview of the survival analysis techniques used. In Section 3.3 the data and the experimental setup are discussed in more detail. The evaluation measures are covered in Section 3.4, followed by the results and discussion in Sections 3.5 and 3.6.

3.2 Survival analysis methods

In survival analysis, one is interested in the timing, T, of a certain event. The survival function can be expressed as the probability of not having experienced the event of interest by some stated time t, hence $S(t) = P(T > t)$. In the context of credit risk, the event of interest is default (together with early repayment and maturity for the mixture cure model with multiple events, see Section 3.2.5). Given the survival function, the probability density function $f(u)$ is given by $f(u) = -\frac{d}{du}S(u)$. Additionally, the hazard function

$$h(t) = \lim_{\tau \to 0} \frac{P(t \leq T < t + \tau \mid T > t)}{\tau}$$

models the instantaneous risk. This function can also be expressed in terms of the survival function and the probability density function

$$h(t) = \frac{f(t)}{S(t)}.$$

In survival analysis, a certain proportion of the cases is censored, which means that for these cases, the event of interest has not yet been observed at the moment of data gathering. In this chapter, we use two different definitions for censoring.

Paper	param./ AFT	Cox PH	AFT/Cox + ext.	non-param.	mixt. cure	multi-event mixt. cure	sample size	number of inputs	evaluation measure
Narain (1992)	X						1242	7	none
Banasik et al. (1999)	X	X					50 000	> 7	Classification
Stepanova and Thomas (2001)		X	X				11 500	16	Classification, AUC, profit measure.
Stepanova and Thomas (2002b)		X	X				50 000	16	Classification, AUC.
Bellotti and Crook (2009)			X				200 000	> 11	Cost of a bad case.
Cao et al. (2009)	X	X		X			25 000	1	AUC
Tong et al. (2012)		X			X		27 527	14	AUC, H-measure, Kolmogorov-Smirnov.
Zhang and Thomas (2012)	X	X	X				27 000	21	Error in default time prediction.
Dirick et al. (2015)					X	X	7521	8	AUC

Table 3.1: *Overview of the existing literature on the use of survival analysis in credit risk modeling. The listed number of inputs is before variable selection (if applicable).*

(1) In the first definition, censored cases are the loans that did not reach their predefined end-date at the moment of data-gathering (called "mature" cases), and did not experience default nor early repayment by this time.

(2) According to the second definition, the uncensored cases are the loans where default has been observed by the end of the observation period. Hence, mature cases and early repayment-cases are labeled as censored, along with the censored cases according to the first definition.

When applying survival analysis to model the time to default, the second definition is used (models in Section 3.2.1–3.2.4). Only for the multiple event mixture cure models in Section 3.2.5, where competing event-types are taken into account, the first definition is used. The censoring indicator for the i-th case is denoted by δ_i, which is equal to 1 for an uncensored observation and is zero when censored.

When using survival models as regression models, a covariate vector and corresponding parameter vector are present. In all models in Section 3.2, the covariate vector is denoted by $\boldsymbol{x}$, and the parameter vector by β.

3.2.1 Accelerated failure time models

Accelerated failure time (AFT) models are fully parametric survival models where explanatory variables act as acceleration factors to speed up or slow down the survival process as compared to the baseline survival function. Formally, this is denoted by

$$S(t;\boldsymbol{x}) = S_0\big(t \cdot \exp(-\beta'\boldsymbol{x})\big)$$

where the event rate is slowed down when $0 < \exp(-\beta'\boldsymbol{x}) < 1$ and speeded up when $\exp(-\beta'\boldsymbol{x}) > 1$. The hazard function is given by

$$h(t;\boldsymbol{x}) = h_0\big(t \cdot \exp(-\beta'\boldsymbol{x})\big)\exp(-\beta'\boldsymbol{x}).$$

In the general form, the accelerated failure time model can be expressed as a log-linear model for the timing of the event of interest $\log(T) = \beta'\boldsymbol{x} + \sigma\epsilon$, with ϵ a random error following some distribution and σ an additional parameter that rescales ϵ. As many classical survival distributions such as the Weibull distribution, exponential distribution and log-logistic distribution have event times that are log-linear, AFT models are often used as a starting point in order to parametrize these distributions. The three models mentioned above are used in the benchmark study and covered further in this section. For a full overview on AFT models and more technical details we refer to Collett (2003) and Kleinbaum and Klein (2011). AFT models are used in the credit risk context by Narain (1992) (who used an exponential distribution), Banasik et al. (1999) (who used exponential and Weibull distributions) and Zhang and Thomas (2012) (who used Weibull, log-logistic and gamma distributions).

Weibull AFT model

The Weibull model in its classical form can be expressed by the following survival and hazard function with scale λ and shape p

$$S(t) = \exp(-\lambda t^p), \quad h(t) = \lambda p t^{p-1}.$$

Using the relationship $\sigma = \frac{1}{p}$, it can be shown that a Weibull-distributed random event time $T_i = \exp(\beta'\boldsymbol{x}_i + \sigma\epsilon_i)$ corresponds to a survival function

$$S_i(t; \boldsymbol{x}) = \exp(-\lambda_i t^{1/\sigma}),$$

where $\lambda_i = \exp\left(-\frac{\beta'\boldsymbol{x}_i}{\sigma}\right)$ is the reparametrization used to incorporate the explanatory variables.

Exponential AFT model

The exponential distribution is a special case of the Weibull distribution, with $p = 1$. This leads to a survival and hazard function

$$S(t) = \exp(-\lambda t), \quad h(t) = \lambda.$$

In the exponential distribution the strong assumption of a constant hazard rate λ is made, and for each case $\lambda_i = \exp\big(-\beta'\boldsymbol{x}_i\big)$. Note that ϵ for the exponential is not rescaled ($\sigma = \frac{1}{p} = 1$).

Log-logistic AFT model

The log-logistic distribution with parameters θ and κ has a survival and hazard function

$$S(t) = \frac{1}{1+\exp(\theta)t^{\kappa}}, \quad h(t) = \frac{\exp(\theta)\kappa t^{\kappa-1}}{1+\exp(\theta)t^{\kappa}}.$$

Using the AFT reparametrization, the relationship $\sigma = \frac{1}{\kappa}$ and the log-logistically distributed event time T_i has a survival function

$$S_i(t;\boldsymbol{x}_i) = \frac{1}{1+\exp(\theta_i)t^{1/\sigma}}$$

where $\theta_i = -\frac{\beta'\boldsymbol{x}_i}{\sigma}$.

3.2.2 Cox proportional hazard model

Another method which is commonly used in survival analysis is the Cox proportional hazards model (see Cox, 1972). This method is more flexible than any AFT model as its baseline hazard function, $h_0(t)$, is a nonparametric one, as opposed to to the parametric baseline hazard function in AFT models. In a Cox proportional hazards model, the hazard function is given by

$$h(t;\boldsymbol{x}) = h_0(t)\exp(\beta'\boldsymbol{x}), \tag{3.1}$$

and the survival function is

$$S(t;\boldsymbol{x}) = \exp\big(-\exp(\beta'\boldsymbol{x})\int_0^t h_0(u)du\big) = \exp\big(-\exp(\beta'\boldsymbol{x})H_0(t)\big),$$

with $H_0(t)$ the cumulative baseline hazard function. In this chapter, Breslow's method is used to estimate the cumulative baseline hazard rate, given

by

$$\widehat{H}_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{r \in R(t_i)} \exp(\beta' \boldsymbol{x}_r)},$$

where $R(t_i)$ denotes the group of individuals at risk at time t_i (which are, in the credit risk context, the ones that have not yet defaulted by time t_i). For more information for the Breslow and other estimators for the Cox PH model, we refer to Klein and Moeschberger (2003). The Cox PH model was first used in the credit context by Banasik et al. (1999).

3.2.3 Cox proportional hazards model with splines

The hazard functions in the Cox PH model, the Weibull AFT model and the exponential AFT model assume a proportional hazards structure with a log-linear model for the covariates. The log-logistic AFT model assumes a proportional odds structure. With these assumptions, hazard or odds ratios are assumed to be constant over time. As a result, for any continuous variable, e.g. age, the default hazard (or odds) ratio between a 25 and a 30-year-old is the same as the hazard (odds) ratio between an 70- and 75-year-old. As it is likely that this assumption does not hold, one has been looking for other functional forms of covariates. For an overview, see Therneau and Grambsch (2000). One of the most popular methods to deal with this is by using splines. Splines are flexible functions defined by piecewise polynomials that are joined in points called "knots". Some constraints are imposed to ensure that the overall curve is smooth. Any continuous variable can be represented by a spline, hence where in formula (3.1) the linear predictor is denoted by

$$\beta' \boldsymbol{x} = \sum_{j=1}^{m} \beta_j x_j;$$

splines can be introduced modeling some or all, say these are $m - l$, continuous covariates by a spline approximation $f_j(\boldsymbol{x}_j)$,

$$\beta' \boldsymbol{x} = \sum_{j=1}^{l} \beta_j x_j + \sum_{j=l+1}^{m} f_j(x_j).$$

For an example, see Figure 3.1. To get a smooth function, a basis of functions with continuous first defivatives is often used to construct a spline function. A popular spline basis is the basis of cubic spline functions

$$1, x, x^2, x^3, (x-\kappa_1)_+^3, \ldots, (x-\kappa_q)_+^3$$

with q knots $\kappa_1, \ldots, \kappa_q$. A spline model is formed by taking a linear combination of the spline basis functions. The disadvantage of power bases, however, lies in the fact that they can become numerically unstable when a large number of knots are included. For this reason, an equivalent basis with more stable numerical properties, the B-spline basis (de Boor, 2001), is nowadays widely used. Both spline models in this study use a cubic B-spline basis in the Cox PH and AFT models. For an overview on splines in a general framework, we refer to Ruppert et al. (2003).

Natural splines

A commonly used modification of the cubic B-spline basis is the natural cubic spline basis. Natural cubic splines satisfy the additional constraint that they are linear in their tails beyond the boundary knots, which are taken to be the endpoints of the data.

Penalized splines

As the number of knots in a spline becomes relatively large, a fitted spline function will show more variation than justified by the data. To limit overfitting, O'Sullivan et al. (1986) introduced a smoothness penalty by integrating the square of the second derivative of the fitted spline function. Later, Eilers et al. (1996) showed that this penalty could also be based on higher-order finite differences of adjacent B-splines. Penalized splines or "P-splines" use the latter method to estimate spline functions.

3.2.4 Mixture cure model

In the medical context, mixture cure models were motivated by the existence of a subgroup of long-term survivors, or a "cured" fraction (see Sy

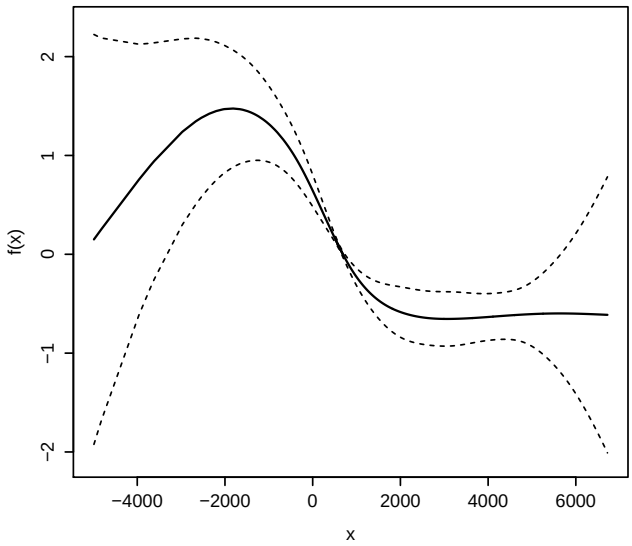

Figure 3.1: *The functional form for one of the covariates x, describing the relationship between x and spline approximation $f(x)$ using penalized splines in a Cox PH model. x is a variable in one of the ten datasets (more details are not disclosed due to confidentiality reasons). The pointwise* 95% *confidence bands are given by the dotted lines.*

and Taylor, 2000a; Peng and Dear, 2000). This subgroup is incorporated in a model through a mixture distribution where a logistic regression model provides a mixing proportion of the "non-susceptible" cases. A survival model describes the cases susceptible to the event of interest.

This type of models is of particular interest in credit risk modeling as the event of interest here, default, will not occur for a very high proportion of the cases. This idea was introduced in the credit risk context for the first time by Tong et al. (2012). In Dirick et al. (2015), a model selection criterion adapted to these models was introduced and applied to credit risk data. For the mixture cure model, the unconditional survival function (for

given values of $\boldsymbol{x}$) is given by:

$$S(t;\boldsymbol{x}) = \pi(\boldsymbol{x})S(t \mid Y = 1, \boldsymbol{x}) + 1 - \pi(\boldsymbol{x}), \tag{3.2}$$

where Y is the susceptibility indicator ($Y = 1$ if an account is susceptible, and $Y = 0$ if not). Note that a new covariate vector $\boldsymbol{x}$ is introduced, which is the covariate vector of the logistic regression model, in this case the binomial logit,

$$\pi(\boldsymbol{x}) = P(Y = 1; \boldsymbol{x}) = \frac{\exp(b'\boldsymbol{x})}{1 + \exp(b'\boldsymbol{x})},$$

with corresponding parameter vector b. In this chapter, the conditional survival function modeling the cases that are susceptible is given by a Cox proportional hazards model,

$$S(t \mid Y = 1, \boldsymbol{x}) = \exp\big(-\exp(\beta'\boldsymbol{x}) \int_0^t h_0(u \mid Y = 1)du\big).$$

Figure 3.2 shows the difference between the survival curves for plain survival functions (such as non-mixture Cox PH and AFT functions) compared to the unconditional survival functions (which are not conditioning on Y, but for given covariate and parameter values) of the mixture cure model. Whereas plain survival curves go to zero as the time goes to infinity, the unconditional survival curves for the mixture cure model "plateau" at a positive value ($1 - \pi(\boldsymbol{x})$).

The mixture cure model is computationally more intensive than plain survival models, as the use of an iterative procedure, the expectation maximization (EM)-algorithm, is needed in order to overcome incomplete information on Y. For more information on mixture cure models, we refer to Farewell (1982), Tong et al. (2012) and Dirick et al. (2015).

3.2.5 Mixture cure model with multiple events

In the medical context it is unusual to ever truly observe cure. In cancer research, for example, a subject might pass away from the specific cancer under research immediately after the observation period, even though

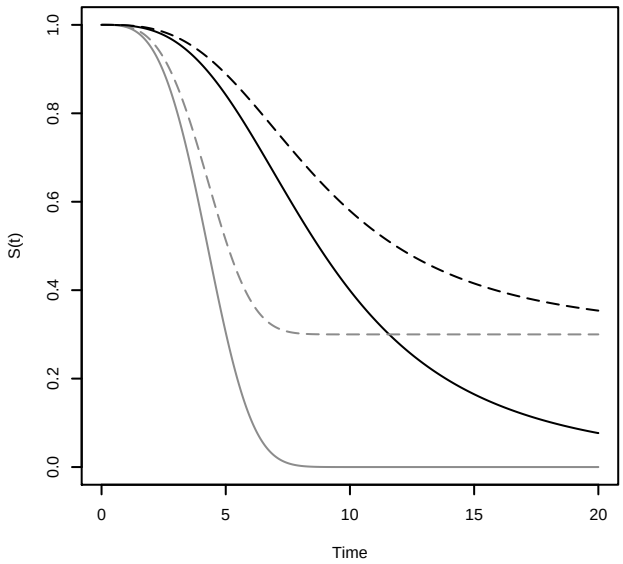

Figure 3.2: *A graphical example pointing out the difference between two plain survival curves and the "unconditional" survival curve in a mixture cure model. Full lines are plain survival curves (modeled using a Weibull AFT model for the gray curve, and a log-logistic AFT model for the black curve), dotted lines represent their corresponding unconditional survival curves in a mixture cure model when assuming a cure rate of 30%.*

having a high probability of being cured. Observed cure does exist in the credit risk context, since as a loan reaches maturity, it is known that default can not occur anymore. As the censoring indicator in the mixture cure model only provides information on whether default took place or not, information on maturity is not used in the model. Another shortcoming is the fact that it does not account for an important "competing risk", early repayment, where a lender repays the loan before the predetermined enddate.

Watkins et al. (2014) recently proposed a method that provides simultaneous modeling of multiple events, along with a mature group. Dirick

et al. (2015) extended this model by allowing for the semi-parametric Cox proportional hazards to model the survival times, instead of the parametric survival models proposed by the former authors. Applied to the credit risk example, three indicators are introduced:

(1) Y_m, indicating that the loan is considered to be mature, so repaid at the indicated end date of the loan;
(2) Y_d, indicating that default takes place;
(3) Y_e, indicating that early repayment takes place.

Note that this set of (Y_m, Y_d, Y_e) is exhaustive and mutually exclusive. However, when an observation is censored (according to the first definition in Section 3.2), it is not known which event type will occur. In analogy to equation (3.2), the unconditional (this is, not conditioning on the Y-triplet) survival function can be written as

$$S(t;\boldsymbol{x}) = \pi_e(\boldsymbol{x}) S_e(t \,|\, Y_e = 1, \boldsymbol{x}) + \pi_d(\boldsymbol{x}) S_d(t \,|\, Y_d = 1, \boldsymbol{x}) + \big(1 - \pi_e(\boldsymbol{x}) - \pi_d(\boldsymbol{x})\big),$$

with $S_e(t \mid Y_e = 1, \boldsymbol{x})$ and $S_d(t \mid Y_d = 1, \boldsymbol{x})$ denoting the conditional survival functions for, respectively, early repayment and default, which are modeled using a Cox proportional hazards model, as in equation (3.2). The $\pi_j(\boldsymbol{x})$'s with $j \in \{e, d\}$ are modeled using a multinomial logit model, hence:

$$\pi_d(\boldsymbol{x}) = P(Y_d = 1; \boldsymbol{x}) = \frac{\exp({b_d}'\boldsymbol{x})}{1 + \exp({b_d}'\boldsymbol{x}) + \exp({b_e}'\boldsymbol{x})}, \tag{3.3}$$

$\pi_e(\boldsymbol{x})$ is found analogously.

3.3 The data and experimental setup

3.3.1 Data preprocessing and missing inputs

Table 3.2 lists the data sets used to evaluate the different survival techniques listed in Section 3.2. We received data sets from five financial institutions in the UK and Belgium, consisting of mainly personal loans

and loans of small enterprises, with varying loan terms. For the banks with data covering several loan terms, the data were split in order to get only one loan term per data set, resulting in ten datasets. Except for bank C, where default is defined as missing two consecutive payments, all banks defined default as missing three consecutive months of payments (Basel II definition).

As some survival analysis techniques are unable to cope with missing data, and with several data sets having a considerable amount of missing inputs, the same pre-processing mechanism to cope with missing data is used for all datasets, as in Dejaeger et al. (2012). For continuous inputs, median imputation is used when $\leq 25\%$ of the values are missing, and the inputs are removed if more than 25% is missing. For categorical inputs, a missing value category is created if more than 15% of the values is missing, otherwise the observations associated with the missing values are removed from the data set.

The number of input variables in the resulting data sets varies from 6 to 31, and the number of observations from 7521 to 80 641. For each observation, an indicator for default, early repayment and maturity is included, taking the value of 1 for the respective event of interest that took place, and 0 for the others (note that only one event type can occur for each observation). Percentages of occurrences of these three event types per data set are given in Table 3.2. For censored observations according to the first censoring definition, all indicators are 0. According to the second censoring definition, only defaults are considered uncensored. In terms of our data sets, this means that censoring rates are ranging from around 20% to 85% according to the first definition (used for the multiple event mixture cure model), whereas censoring percentages are not lower than 94.56% up to 98.16% according to the second definition.

Additionally, a time variable is given to each observation, representing the respective month of the event, which takes an integer value. Note that the time variable for a mature event is always equal to the length of the loan term (e.g. a mature loan for data set 5 has value 24), and the time variable for a censored event is given by the last observed month in which

a repayment was observed to take place.

Data set	Bank	loan term (months)	size	inputs (nmbr)	Cat. (nmbr)	Cont. (nmbr)	default (%)	early (%)	mature (%)
DS 1	A	36	42 903	31	13	18	4.03	26.80	49.72
DS 2	A	48	46 970	31	13	18	4.02	28.61	40.49
DS 3	A	60	80 641	31	13	18	5.44	32.74	24.80
DS 4	B	12	10 027	13	7	6	2.73	53.80	24.43
DS 5	B	24	9979	13	7	6	4.74	38.46	28.88
DS 6	B	36	7521	13	7	6	5.00	39.78	3.58
DS 7	C	48	9980	6	5	1	1.84	9.20	19.80
DS 8	C	60	17 378	6	5	1	1.84	8.95	4.13
DS 9	D	37	35 856	11	8	3	3.56	19.27	46.83
DS 10	E	60	9785	8	4	4	1.62	10.09	17.85

Table 3.2: *Data set specifications.*

3.3.2 Experimental setup

Each data set was randomly split up in a training set and a test set consisting of 2/3 and 1/3 of the observations respectively. The models are induced on the training sets, and the corresponding test sets are used for evaluation. Several random splits of the data sets were initially tested to ensure robustness of the reported results.

For all models, the software R is used. AFT and Cox proportional hazards modeling is possible through the use of the R-package `survival` (Therneau, 2014), with additional use of functions `ns` and `pspline` for inclusion of natural splines and penalized splines in the covariates respectively. An ad hoc method was used to decide on which of the continuous variables a spline function should be introduced. Using the `pspline`-function on each continuous variable in the model seperately, the resulting spline curves were inspected to track some possible non-linear relationships, with knots determined by the adapted AIC method (Eilers et al. (1996), included in the package). The resulting Cox proportional hazards and accelerated failure time pspline models consisted of all the P-splines where non-linear relationships were observed. As the `ns`-function does not have a built-in

function to optimize the number of knots, the same continuous variables and number of knots were chosen as in the pspline models. For some of the data sets, the number of splines or knots using the natural splines were altered in comparison with the pspline models, in order to get a feasible fit.

For the mixture cure model, the R-package `smcure` by Cai et al. (2012a,b) is used. An extended code based on this package, as in Dirick et al. (2015), is used for the multiple event mixture cure model (code available upon request).

3.4 Performance criteria/evaluation metrics

To evaluate performance of the survival models, three main performance criteria were used: the area under the curve of the receiver operating characteristics curve, default time prediction and an annuity theory-based measure. Note that other statistical evaluation measures for survival analysis (e. g. martingale and deviance residuals) exist, but in this study, priority was given to evaluation methods that are less abstract from a banking point of view.

3.4.1 AUC in ROC-curves

In the credit risk context, an ubiquitous method to evaluate binary classifiers is by means of the receiver operating characteristics curve. This curve illustrates the performance of a binary classifier for each possible threshold value, by plotting the true positive rate against the false positive rate. The specific measure of interest is the area under the curve (AUC), which can also be computed in the context of survival analysis. In this context, evaluation is possible at any timepoint of the survival curve (see Heagerty and Saha, 2000). For each data set and each model, the AUC for the testsets at 1/3 and 2/3 of the time to maturity, and at the maturity time itself (which is equal to the loan term) is listed in Table 3.3.

Despite the fact that AUC and other classification-based evaluation

methods are most common in the literature (see Table 3.1), this way of evaluating survival analysis in credit risk has some disadvantages and does not fully highlight the benefits of using survival this context. First of all, as using AUC in the survival analysis requires taking a point in time at which we evaluate, the results are strongly dependent of the chosen evaluation times. Because only some points in time are chosen, the AUC does not fully summarize the time aspects of survival analysis. Secondly, the financial aspect is neglected. The next two sections focus on the timing aspects and economic/financial evaluation respectively.

3.4.2 Evaluation through default time prediction

When evaluating through default time prediction, we look at how we are able to predict the default times of the defaults in the testset. A survival curve does not give one time estimate, but a distribution of time estimates. With a high amount of censoring, mean values of these survival analysis do not give good predictors. Zhang and Thomas (2012) use parametric survival models to model recovery rates, and use quantiles of these models to minimise the MSE and MAE. They do this by looking at each percentile of the fitted survival model using the training set only, and by calculating the squared and absolute deviations from the predictions to the real values of the default cases. Next, the percentiles resulting in the lowest deviations are withheld and used to compute the deviations in the test set.

We use the same method as Zhang and Thomas (2012), but consider the default time instead of recovery values, and look at each permille. The results are listed in Table 3.4, where the MSE-columns list the mean of the squared differences between the predicted and real default times, and the MAE-columns list the mean of the absolute differences between the predicted and real default times.

Note that the part of the data set which is evaluated is considerably smaller here compared to the AUC method. A schematic representation is given in Figure 3.3. Each letter represents an observation in the entire data set, where four possible end states are possible: Early repayment

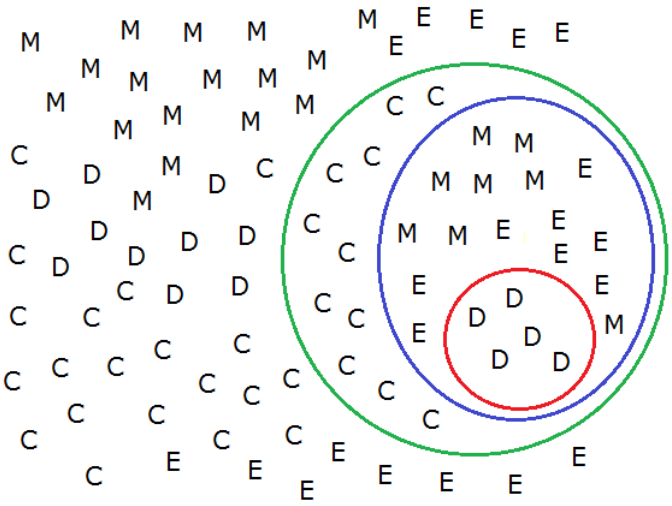

Figure 3.3: *Schematic representation of the data set. Each letter represents an observation in the data set. The data set elements that are in the test set are in the largest green circle. All test set elements are evaluated using the AUC evaluation method. The uncensored test set elements that are in the middle (blue) circle are evaluated through the economic evaluation method using annuity theory. Default time prediction evaluation can only be performed on the defaulted elements of the test set, encompassed by the smallest (red) circle.*

("E"), Default ("D"), Maturity ("M") and Censored ("C"). The green circle encompasses the test set elements, which are all evaluated when computing the AUC. The default time prediction method, however, only evaluates the default times of the "actual" defaults in the test set, which are in the red circle.

3.4.3 Evaluation using annuity theory

When banks grant a loan to a customer, they are particularly interested in the expected future value at the end of the loan term. One can use the principles of annuity theory (for an overview, see Kellison and Irwin, 1991) to compute this value, though these basic principles do not incorporate risk,

hence these formulas start from the assumption that loans will be repaid with a 100% certainty. Including this risk aspect is exactly what can be done using survival analysis, as it provides us with an accurate estimate for the probability that a customer is still repaying his loan at every time instant of the survival curve. The idea of using expected profit measures when using survival analysis was first introduced by Thomas (2009, Section 4.6). In this section, we use similar formulas that are also appropriate in the mixture cure context.

In this study, we computed the true future value of the uncensored test set loans (given by the observations in the blue circle in Figure 3.3), taking into account their true end-state (default, early repayment or maturity), and compare them to their estimated values using each of the survival models. In order to make the results comparable, some assumptions are made and applied when evaluating the models for all data sets. All existing formulas can be extended to account for deviations from these assumptions (early repayment penalties etc.).

(1) loans are repaid at the end of each month, with a fixed sum.

(2) The (yearly) interest rate i_y used is 5%.

(3) The loans are treated as if they all started at the same point in time, in order to make them comparable.

(4) Loss given default is set to 1, meaning that nothing of the remaining debt is assumed to be recovered when default occurs.

(5) There is assumed to be no penalty charge in case of early repayment.

Let us introduce:

(a) $\boldsymbol{L_s}$ the initial amount of the loan, or the debt of subject s;

(b) $\boldsymbol{R_s}$ the fixed sum of the monthly repayment for subject s;

(c) n the number of periods;

(d) i the monthly interest rate ($i = (1 + i_y)^{1/12} - 1$);

(e) $(E)FV$ the (expected) future value of a loan.

The fixed sum $\boldsymbol{R_s}$ consists of a repayment of the loan, $a_{s,j}$, and some interest paid, $p_{s,j}$, each in month j. Hence, $\boldsymbol{R_s} = a_{s,j} + p_{s,j}$. Note that where $\boldsymbol{R_s}$ remains constant, $a_{s,j}$ and $p_{s,j}$ change over time, where $a_{s,j}$ increases and $p_{s,j}$ decreases. Then,

$$\boldsymbol{L_s} = \sum_{j=1}^{n} a_{s,j} = a_{s,1} \cdot \sum_{j=1}^{n} (1+i)^{j-1}.$$

Hence it can be shown that

$$\boldsymbol{R_s} = \frac{i}{1-(1+i)^{-n}} \boldsymbol{L_s}.$$

A bank can reinvest the repayment sums $\boldsymbol{R_s}$ as soon as they are paid by the client. Assume that the same interest rate applies. If there is no risk for default nor early repayment, the future value can be given by

$$FV_s = \boldsymbol{R_s}\big((1+i)^{n-1} + (1+i)^{n-2} + \ldots + (1+i)^0\big) = \boldsymbol{R_s}\frac{(1+i)^n - 1}{i}. \quad (3.4)$$

For the uncensored test set loans, we wish to estimate the future loan values. In Section 3.4.3, we describe how we compute the true future values when knowing the eventual state ("D", "M" or "E"), in Section 3.4.3, the expected future loan value is estimated when using the model predictions. Table 3.5 lists the mean absolute differences between the real future values and the expected future loan values using the model estimations. In table 3.6, we look at the mean expected future values per loan and compare them with the mean real future value.

The true future loan values

The true future loan value depends on the eventual loan outcome or state. For mature loans, Equation (3.4) can be used with n the total number of periods or the loan term. Hence

$$FV_{s \in \text{mature}} = \boldsymbol{R_s}\frac{(1+i)^n - 1}{i}.$$

For the future value of a loan with early repayment, the resulting amount of the debt in any time period k is given by

$$\boldsymbol{L_{s,k}} = \Big(1 - \frac{(1+i)^k - 1}{(1+i)^n - 1}\Big)\boldsymbol{L_s}.$$

When an early repayment takes place in period k, we assume that the loan is repaid as usual until this period k, and that the sum $\boldsymbol{L_k}$ is fully being repaid in this period. This sum can still be reinvested for $n-k-1$ periods. Note that early repayment always yields a smaller revenue compared to a mature loan,

$$FV_{s\in\text{early}} = \boldsymbol{R_s}\Big(\sum_{j=1}^{k}(1+i)^{n-j}\Big) + \boldsymbol{L_{s,k}}(1+i)^{n-k-1}.$$

The future value for a loan where defaults take place after k months is equal to

$$FV_{s\in\text{default}} = \boldsymbol{R_s}\Big(\sum_{j=1}^{k}(1+i)^{n-j}\Big),$$

hence we assume that when default takes place, nothing of the remaining sum $\boldsymbol{L_k}$ is recovered.

The expected future loan values using the model predictions

In each of the survival models in Section 3.2.1–3.2.3, the model provides us with a survival probability estimate at each point in time. We denote $\widehat{S}(t)^d_{s,m}$ the estimated probability that subject s has not defaulted by time t, using model m. Then we can calculate the expected terminal value of a loan according to a certain model m as follows:

$$EFV_{s,m} = \boldsymbol{R_s}\Big(\sum_{j=1}^{n}\widehat{S}(j)^d_{s,m}(1+i)^{n-j}\Big).$$

For the mixture cure model in Section 3.2.4 , we also have probabilities of being susceptible to default or not (PD and $1-PD$) for every subject.

We define

$$PD_s = \hat{\pi}(\boldsymbol{x_s}) = \frac{\exp(\hat{b}'\boldsymbol{x_s})}{1+\exp(\hat{b}'\boldsymbol{x_s})}, \tag{3.5}$$

then we have

$$\begin{aligned} EFV_{s,m} &= PD_s \cdot \boldsymbol{R_s}\Big(\sum_{j=1}^{n} \widehat{S}(j)^d_{s,m}(1+i)^{n-j}\Big) \\ &\quad +(1-PD_s)\cdot \boldsymbol{R_s}\frac{(1+i)^n-1}{i}. \end{aligned}$$

For the multiple event mixture cure model in Section 3.2.5, we have in addition probabilities of early repayment PE and probabilities of maturity PM. Working in a similar fashion as in (3.5) but using the multinomial logit (3.3). Additionally, $\widehat{S}(t)^e_{s,m}$ is the estimated probability that subject s has not repaid early by time t for every subject. The expected future value is given by

$$\begin{aligned} EFV_{s,m} &= PD_s \cdot \boldsymbol{R_s}\Big(\sum_{j=1}^{n} \widehat{S}(j)^d_{s,m}(1+i)^{n-j}\Big) \\ &\quad +PM_s \cdot \boldsymbol{R_s}\frac{(1+i)^n-1}{i} \\ &\quad +PE_s\cdot\Bigg(\boldsymbol{R_s}\Big(\sum_{j=1}^{n} \widehat{S}(j)^e_{s,m}(1+i)^{n-j}\Big) \\ &\quad +\sum_{j=1}^{n-1}\Big(\big(\widehat{S}(j-1)^e_{s,m}-\widehat{S}(j)^e_{s,m}\big)\boldsymbol{L_{s,j}}(1+i)^{n-j-1}\Big)\Bigg). \end{aligned}$$

Evaluating the expected future value with respect to the real future value.

For each of the uncensored test set cases, the real future value can be computed giving the eventual outcome, and be compared with the expected future values using the models. Table 3.5 lists the mean of the absolute differences between the expected and the real values per case. In Table 3.6,

	Data set 1			Data set 2			Data set 3			Data set 4			Data set 5		
Method \ AUC	1/3	2/3	3/3	1/3	2/3	3/3	1/3	2/3	3/3	1/3	2/3	3/3	1/3	2/3	3/3
AFT Weibull	0.828	0.826	0.847	0.831	0.835	0.827	0.812	0.809	0.816	0.846	0.804	0.779	0.704	0.711	0.716
AFT exponential	0.829	0.826	0.847	0.831	0.834	0.826	0.812	0.809	0.815	0.845	0.804	0.779	0.703	0.710	0.716
AFT loglogistic	0.829	0.826	0.847	0.831	0.835	0.827	0.812	0.809	0.816	0.846	0.804	0.779	0.704	0.711	0.716
AFT Weib. w nat. splines	0.832	0.829	0.851	0.836	0.837	0.830	0.819	0.817	0.820	0.829	0.792	0.770	0.670	0.680	0.674
AFT expo w nat. splines	0.834	0.829	0.848	0.837	0.837	0.830	0.819	0.816	0.819	0.796	0.804	0.776	0.668	0.684	0.680
AFT loglog w nat. splines	0.831	0.829	0.849	0.837	0.839	0.831	0.828	0.820	0.812	0.796	0.793	0.770	0.669	0.684	0.679
AFT Weib. w penal. splines	0.834	0.831	0.852	0.835	0.839	0.831	0.819	0.816	0.820	0.844	0.805	0.779	0.698	0.705	0.711
AFT expo w penal. splines	0.835	0.830	0.850	0.837	0.838	0.830	0.819	0.815	0.819	0.842	0.822	0.790	0.698	0.707	0.711
AFT loglog w penal. splines	0.835	0.831	0.852	0.837	0.839	0.832	0.819	0.816	0.820	0.843	0.821	0.790	0.698	0.706	0.710
Cox PH	0.828	0.824	0.846	0.830	0.834	0.826	0.812	0.809	0.815	0.846	0.804	0.780	0.704	0.711	0.716
Cox PH w nat. splines	0.852	0.837	0.854	0.854	0.851	0.835	0.829	0.821	0.810	0.785	0.779	0.757	0.677	0.695	0.698
Cox PH w penal. splines	0.832	0.827	0.849	0.838	0.838	0.830	0.820	0.817	0.821	0.834	0.820	0.789	0.684	0.691	0.698
Mixt. cure	0.829	0.828	0.847	0.827	0.833	0.825	0.817	0.814	0.823	0.875	0.833	0.787	0.702	0.705	0.699
Multi-event mixt. cure	0.829	0.824	0.844	0.832	0.835	0.825	0.811	0.806	0.816	0.821	0.801	0.779	0.715	0.716	0.715

Table 3.3: *Test set "Areas Under the Curve" (AUC) for the different methods applied to the 10 data sets when evaluating at several timepoints, corresponding to* $1/3, 2/3$ *and the full loan term, which depends on the data set. The three best values are underlined. AUCs at the three timepoint are comparable within one dataset, so columnwise.*

	Data set 6			Data set 7			Data set 8			Data set 9			Data set 10		
Method\ AUC	1/3	2/3	3/3	1/3	2/3	3/3	1/3	2/3	3/3	1/3	2/3	3/3	1/3	2/3	3/3
AFT Weibull	0.736	0.706	0.665	0.749	0.715	0.654	0.596	0.668	0.662	0.852	0.850	0.849	0.711	0.751	0.766
AFT exponential	0.736	0.706	0.664	0.745	0.714	0.653	0.598	0.667	0.659	0.852	0.849	0.849	0.710	0.750	0.764
AFT loglogistic	0.736	0.706	0.664	0.751	0.715	0.654	0.599	0.669	0.663	0.852	0.850	0.849	0.712	0.752	0.766
AFT Weib. w nat. splines	0.717	0.705	0.658	0.741	0.692	0.615	0.616	0.664	0.652	0.861	0.857	0.858	0.741	0.779	0.799
AFT expo w nat. splines	0.716	0.706	0.659	0.738	0.690	0.613	0.617	0.665	0.647	0.861	0.857	0.857	0.725	0.770	0.785
AFT loglog w nat. splines	0.716	0.704	0.658	0.741	0.692	0.616	0.617	0.664	0.653	0.862	0.858	0.858	0.703	0.747	0.772
AFT Weib. w penal. splines	0.719	0.711	0.650	0.500	0.500	0.500	0.596	0.668	0.662	0.860	0.853	0.852	0.725	0.785	0.795
AFT expo w penal. splines	0.720	0.711	0.650	0.745	0.714	0.653	0.598	0.667	0.659	0.860	0.853	0.851	0.722	0.778	0.787
AFT loglog w penal. splines	0.719	0.711	0.649	0.500	0.500	0.500	0.599	0.669	0.663	0.860	0.854	0.853	0.710	0.752	0.769
Cox PH	0.736	0.705	0.663	0.733	0.710	0.652	0.596	0.668	0.663	0.852	0.850	0.849	0.712	0.750	0.765
Cox PH w nat. splines	0.708	0.698	0.656	0.731	0.687	0.610	0.582	0.664	0.649	0.859	0.858	0.857	0.738	0.778	0.797
Cox PH w penal. splines	0.719	0.711	0.649	0.732	0.710	0.651	0.602	0.670	0.664	0.859	0.856	0.855	0.723	0.782	0.792
Mixt. cure	0.727	0.702	0.674	0.654	0.657	0.623	0.630	0.682	0.643	0.854	0.852	0.850	0.699	0.739	0.748
Multi-event mixt. cure	0.729	0.703	0.665	0.723	0.706	0.654	0.630	0.678	0.640	0.850	0.849	0.849	0.695	0.750	0.760

Table 3.3 (continued)

the mean expected future values of all uncensored test set loans are listed, and can be compared with the mean of the true future loan value at the bottom of the table.

3.5 Results

The results in Tables 3.3– 3.6 are grouped per evaluation measure. For Tables 3.4 and 3.5, we used a notational convention where the best test result (each time the smallest value) per data set is underlined and denoted in bold face. Performances that are significantly different at a 5% level from the top performance with respect to a one-sided Mann - Whitney test are denoted in bold face (a Bonferroni-correction was used due to multiple testing). As the AUC-values in Table 3.3 are point estimates and do not represent samples, here simply the three highest values are underlined for each evaluation time and dataset. In Table 3.6 the three values that lie closest to the mean future value per loan are underlined. Table 3.7 summarizes the results of all preceding tables by giving the average ranks of the models for all evaluation methods.

In Table 3.3 we note that sample size is of real importance to get better receiver operating characteristics curves, as AUC-values are consistently larger for datasets with more observations. However, it is hard to draw conclusions regarding the preferred survival method when looking at the AUC alone, as the values are very close to each other (we note that ties in Table 3.3 are due to rounding). This can also be seen in Table 3.7, as average rankings range from 5.30 to 9.60 (compared to $[2.6, 12.7]$, $[2.2, 13.4]$, $[3.10, 13.7]$ and $[3.3, 11.1]$ in the other categories). Because of this, preferred methods seem to be very different from one data set to another. The first three columns of Table 3.7 do seem to point out that there is a tendency towards both AFT and Cox PH methods where penalized splines are included.

In Table 3.4, quite a few performance measures are not significantly different from the top performance at the 5% level. However, an overall trend for these evaluation measures is that the exponential AFT models

(both with splines and without) seem to have default time predictions that are significantly far off the true default times. For most data sets, the inferiority of all AFT models in comparison with Cox PH-based models becomes apparent when looking at Tables 3.4 and 3.7. With average ranks between 2.2 and 4, it seems that the default time prediction measure clearly favors the plain Cox PH model, the Cox PH model with penalized splines and the mixture cure model.

Table 3.5 lists the mean of the absolute differences between the model expected future loan value estimates and the true values. Note that these differences are bigger for loans with a longer loan term, which is logical, as here the loan amounts are larger too. On an individual level, the differences can get to a substantial size (for example, in dataset 3), but considering Table 3.6 we note that the mean expected values per loan are close to the mean real value of the loans for all methods. These results clearly highlight the abilities of survival analysis in the credit risk context. Consulting Table 3.7, we see that the results for Table 3.5 and 3.6 differ quite heavily, although both using annuity theory as an evaluation measure. The explanation lies again in the fact that all results lie quite close to each other. The results from the Mann-Whitney test in Table 3.5 show again that exponential AFT models are inferior to other models. Additionally, the multiple event mixture cure model leads to significantly inferior results. Surprisingly, AFT exponential models do seem to perform well when comparing expected future value per loan with the true future value (Table 3.6), though the single event mixture cure model is performing slightly better, being among the best three models seven out of ten times. The exponential AFT models, however, seem to be dominating in the seventh column of Table 3.7, though being clearly inferior for all other evaluation measures.

To get an overall idea of the model performance, the last column of Table 3.7 gives the average model rank over all evaluation methods. Cox PH-based models (in particular, Cox PH with penalized splines and the mixture cure model) seem to outperform AFT-based models, with the exception of some AFT models with penalized splines. An important ob-

	Data set 1		Data set 2		Data set 3		Data set 4		Data set 5	
Method \ deviation measure	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
AFT Weibull	**333.43**	13.51	**662.40**	18.39	**612.89**	18.16	121.34	4.29	79.48	6.75
AFT exponential	**434.42**	**15.94**	**762.30**	**20.67**	**677.47**	**19.40**	**4233.42**	**20.17**	**127.89**	8.46
AFT loglogistic	**344.53**	13.55	678.44	18.36	**613.16**	18.22	123.08	4.32	79.52	6.75
AFT Weibull w nat. splines	**321.05**	13.34	929.93	18.70	585.33	18.09	206.69	5.07	81.17	6.88
AFT exponential w nat. splines	**414.85**	**15.74**	**748.26**	**20.85**	**654.59**	**19.34**	**20028.87**	**28.21**	137.32	8.67
AFT loglogistic w nat. splines	**319.60**	13.40	544.29	17.67	**676.65**	18.63	265.64	5.36	82.12	6.94
AFT Weibull w penal. splines	**319.88**	13.26	676.33	18.12	579.55	17.88	120.07	4.35	74.74	6.68
AFT exponential w penal. splines	**415.53**	**15.70**	**770.73**	**20.50**	**652.47**	**19.22**	**6302.19**	**19.45**	115.23	8.26
AFT loglogistic w penal. splines	**328.10**	13.28	574.91	17.79	573.73	17.93	112.45	4.27	74.08	6.69
Cox PH	229.08	11.91	412.77	15.62	510.87	16.93	12.59	2.97	63.78	6.34
Cox PH w nat. splines	235.98	12.03	415.69	15.71	532.30	17.41	17.02	3.25	71.60	6.83
Cox PH w penal. splines	235.61	12.14	412.00	15.69	512.63	17.06	14.44	3.12	70.90	6.80
Mixture cure	233.33	11.99	412.21	15.68	527.01	17.22	12.55	2.86	68.54	6.63
Multiple event mixture cure	262.81	12.65	475.08	16.72	564.77	17.70	13.47	3.11	68.11	6.54

Table 3.4: *Deviation measures when predicting the default times for observed defaults in the test set of the 10 datasets, using different methods. Top performances for each test set are underlined. Performances that are significantly different at a 5% level from the top performance with respect to a one-sided Mann - Whitney test are denoted in bold face.*

	Data set 6		Data set 7		Data set 8		Data set 9		Data set 10	
Method \ deviation measure	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
AFT Weibull	88.29	6.87	281.89	12.50	434.65	16.52	**385.94**	13.18	490.57	18.37
AFT exponential	**136.93**	8.14	444.42	17.81	**768.92**	20.76	**487.06**	**16.32**	837.56	24.31
AFT loglogistic	90.79	6.89	263.62	12.53	436.83	16.55	**382.82**	13.13	494.64	18.44
AFT Weibull w nat. splines	93.67	7.30	271.07	12.07	445.57	17.04	**294.80**	12.52	477.57	17.97
AFT exponential w nat. splines	**137.05**	8.83	**403.81**	17.10	888.38	21.88	**374.03**	**15.25**	951.93	24.10
AFT loglogistic w nat. splines	94.92	7.36	254.16	12.13	449.19	17.16	**293.16**	12.48	504.05	18.90
AFT Weibull w penal. splines	97.36	7.40	147.03	10.26	434.65	16.52	270.87	12.07	462.79	17.89
AFT exponential w penal. splines	**141.15**	9.03	444.42	17.81	**768.92**	20.76	**351.84**	**14.61**	864.66	23.82
AFT loglogistic w penal. splines	99.00	7.47	146.94	10.26	436.83	16.55	265.12	12.08	491.68	18.53
Cox PH	89.09	7.15	245.18	11.25	414.76	16.34	185.27	10.37	505.24	18.31
Cox PH w nat. splines	94.34	7.16	285.44	12.75	415.78	16.62	189.38	10.42	521.93	18.20
Cox PH w penal. splines	92.15	7.13	237.89	11.32	413.55	16.25	188.13	10.40	491.15	17.56
Mixture cure	92.28	7.24	359.54	15.23	410.42	16.37	186.75	10.62	511.81	18.00
Multiple event mixture cure	90.17	7.74	278.32	12.19	418.88	16.48	210.38	11.06	524.63	18.86

Table 3.4 (continued)

servation seems to be that the multiple event mixture cure model seems clearly inferior to most other methods.

3.6 Discussion

In this chapter, we studied the performance of several survival analysis techniques in the credit scoring context. Ten real-life data sets were used, and we used three main evaluation measures to assess model performance: AUC, default time prediction differences, and future loan value estimation. It is shown that Cox PH-based models all work particularly well, especially a Cox PH model in combination with penalized splines for the continuous covariates. Where this model generally outperforms the multiple event mixture cure model, the mixture cure model does not perform significantly different in most of the cases, and is among the top models. AFT exponential models (with and without splines) are clearly inferior to other models, and overall, penalized splines outperform natural splines.

Starting from these findings, it would be interesting to further extend the mixture cure model and study the performance of the resulting model in comparison with a Cox PH model with penalized splines. This could be done by allowing for splines in the continuous covariates. This study also points out that finding an appropriate evaluation measure to compare survival analysis remains an interesting challenge, as the AUC does not seem to have the right properties to really distinguish one method from another.

MAD from FV	DS 1	DS 2	DS 3	DS 4	DS 5	DS 6	DS 7	DS 8	DS 9	DS 10
AFT Weibull	**334.1**	748.8	1695.7	29.3	133.2	384.0	669.0	1522.4	420.1	887.6
AFT exponential	**350.4**	**769.2**	**1709.6**	**32.6**	**139.1**	386.7	687.2	1524.3	**446.6**	**917.1**
AFT loglogistic	**334.5**	750.6	1699.7	29.4	133.4	384.9	670.1	1524.2	424.7	889.3
AFT Weibull w nat. splines	333.3	746.6	1678.8	29.3	133.7	389.2	663.0	1551.3	403.8	884.9
AFT exponential w nat. splines	**350.1**	**766.8**	1694.2	**32.6**	**139.9**	391.4	681.6	1537.1	**419.6**	917.2
AFT loglogistic w nat. splines	333.9	748.0	1708.8	29.3	134.0	390.3	664.2	1547.6	404.9	885.7
AFT Weibull w penal. splines	333.4	746.4	1680.4	29.3	132.5	388.5	645.7	1522.4	403.7	884.3
AFT exponential w penal. splines	**350.0**	**767.6**	1695.6	**32.5**	**138.7**	390.9	687.2	1524.3	**419.8**	919.0
AFT loglogistic w penal. splines	333.9	748.5	1683.3	29.2	132.7	389.2	645.8	1524.2	404.5	887.5
Cox PH	**332.0**	744.4	1693.1	**30.2**	**135.9**	383.2	667.1	1519.7	**428.0**	887.5
Cox PH w nat. splines	**335.2**	756.9	1716.1	30.0	136.2	388.5	661.5	1556.7	409.1	885.4
Cox PH w penal.	**332.0**	740.3	1673.7	29.9	134.5	387.6	659.3	1519.7	409.1	888.6
Mixture cure	330.8	746.2	1692.8	29.5	135.7	382.0	672.4	1526.4	411.5	906.3
Multiple event mixture cure	**409.9**	**940.9**	**2109.5**	**31.5**	**148.9**	**435.3**	**744.5**	**1691.1**	**467.4**	**1054.6**

Table 3.5: *Analyzing model performance using financial metrics. Mean absolute deviations from the real future loan values for the uncensored cases (first definition) of the test set. Top performances for each test set are underlined. Performances that are significantly different at a 5% level from the top performance with respect to a one-sided Mann - Whitney test are denoted in bold face.*

Mean EFV per loan	DS 1	DS 2	DS 3	DS 4	DS 5	DS 6	DS 7	DS 8	DS 9	DS 10
AFT Weibull	8180.1	14 213.5	19 585.4	1020.8	2119.1	4067.2	15 793.6	21 021.2	9686.8	21 972.0
AFT exponential	8159.9	14 186.5	19 566.7	1016.9	2111.9	4063.7	15772.6	21 017.0	9655.7	21 935.4
AFT loglogistic	8179.5	14 210.9	19 580.4	1020.7	2118.8	4065.9	15 792.2	21 018.5	9681.2	21 969.9
AFT Weibull w nat. splines	8178.9	14 211.0	19 586.2	1020.7	2119.0	4061.4	15 803.0	20997.0	9700.6	21 968.5
AFT exponential w nat. splines	8158.5	14183.7	19 567.1	1016.8	2111.5	4058.5	15 782.0	21 008.5	9680.1	21929.6
AFT loglogistic w nat. splines	8177.8	14 207.4	19534.0	1020.6	2118.6	4060.0	15 801.7	20 999.8	9698.1	21 975.6
AFT Weibull w penal. splines	8179.2	14 212.6	19 588.2	1020.8	2119.9	4060.9	15 830.3	21 021.2	9703.0	21 971.1
AFT exponential w penal. splines	8158.8	14183.3	19 568.2	1016.9	2112.4	4057.9	15772.6	21 017.0	9682.3	21927.8
AFT loglogistic w penal. splines	8178.5	14 207.7	19 583.3	1020.7	2119.7	4060.0	15 830.2	21 018.5	9700.9	21 972.1
Cox PH	8183.9	14 222.2	19 596.5	1019.8	2115.7	4068.6	15 796.6	21 025.3	9677.8	21 971.9
Cox PH w nat. splines	8175.6	14182.3	19515.8	1019.7	2115.7	4062.8	15 805.7	20990.2	9695.9	21 967.7
Cox PH w penal. splines	8182.2	14 214.9	19 590.6	1019.8	2117.7	4062.5	15 804.7	21 025.5	9697.0	21 966.1
Mixture cure	8099.1	14 004.2	19120.2	1018.2	2101.3	4007.3	15708.1	20803.8	9633.0	21779.4
Multiple event mixture cure	8185.0	14 217.4	19 588.4	1020.3	2115.8	4070.3	15 794.5	21 011.6	9693.7	21 951.2
Mean FV per loan	*8164.3*	*14173.2*	*19339.2*	*1012.2*	*2096.5*	*3966.0*	*15590.8*	*20137.1*	*9649.7*	*21563.7*

Table 3.6: *Analyzing model performance using financial metrics. Mean expected future loan values of the uncensored cases (first definition) of the test set. The three best values are underlined.*

	AUC 1/3	AUC 2/3	AUC 3/3	MSE	MAE	MAD from FV	EFV vs FV	**ALL**
AFT Weibull	7.50	7.60	7.50	7.60	7.50	6.30	10.30	7.76
AFT exponential	8.10	9.10	8.90	13.00	13.40	11.30	4.00	9.69
AFT loglogistic	7.10	7.40	6.70	8.60	8.20	7.60	8.50	7.73
AFT Weibull w nat. splines	6.50	8.10	7.10	8.50	8.10	5.20	8.70	7.46
AFT exponential w nat. splines	7.40	8.00	8.60	12.70	13.30	11.00	3.30	9.19
AFT loglogistic w nat. splines	7.40	8.00	8.30	8.70	9.20	6.70	7.30	7.94
AFT Weibull w penal. splines	7.65	5.65	6.45	6.30	5.70	3.10	11.10	6.56
AFT exponential w penal. splines	6.10	5.30	6.20	12.40	12.30	10.80	3.60	8.10
AFT loglogistic w penal. splines	7.05	5.45	6.45	6.50	6.60	4.60	9.50	6.59
Cox PH	8.60	9.40	8.00	2.60	2.20	5.80	9.90	6.64
Cox PH w nat. splines	7.50	7.20	7.90	5.80	5.70	8.50	5.50	6.87
Cox PH w penal. splines	6.80	5.20	5.80	3.10	3.20	4.30	9.90	5.47
Mixture cure	8.20	9.00	8.40	4.00	3.90	6.10	4.20	6.26
Multiple event mixture cure	9.10	9.60	8.70	5.20	5.70	13.70	9.20	8.74

Table 3.7: *Average ranking of the models used depending on the evaluation method. The three best values are underlined. The last column is the average ranking over all evaluation methods.*

Chapter 4

Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models

Abstract

The prediction of the time of default in a credit risk setting via survival analysis needs to take a high censoring rate into account. This rate is due to the fact that default does not occur for the majority of debtors. Mixture cure models allow to model the part of the loan population that is insusceptible to default. In this chapter we extend the mixture model to include time-varying covariates. We illustrate the method via simulations and by incorporating macro-economic factors as predictors for an actual bank data set.

This chapter is based on Dirick, L., Bellotti, T., Claeskens, G. and Baesens, B. (2015). Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. Working paper, submitted.

4.1 Introduction

With recent compliance guidelines such as the Basel accords, increased attention is devoted to more accurate calculations of the minimum amount of capital banks need to hold to provide a buffer against unexpected losses (Van Gestel and Baesens, 2008). Where the probability of default (PD) of a certain loan applicant is usually constructed using classification techniques such as logistic regression, other methods gained more importance. Survival analysis is an interesting tool here as this method enables modeling of time until default, and not just whether a certain customer will default.

With important applications in actuarial sciences through lifetables and currently largely used in medical science (see Collett, 2003; Cox and Oakes, 1984), survival analysis was first introduced in the credit scoring context by Narain (1992). While initially using fully parametric accelerated failure time survival models, other authors extended the idea of Narain (1992), using a Cox proportional hazards (PH) model (see Banasik et al., 1999), extensions on Cox PH models (Stepanova and Thomas, 2002b) and including macro-economic variables (MVs) through time-varying covariates (TVCs) in Cox PH models (see Bellotti and Crook, 2009). In these papers, it is shown that survival analysis is a competitive method to logistic regression, and extending the Cox PH model further improves the accuracy of the estimated PD.

The survival function in standard survival analysis is given by $S(t) = P(T > t)$, which is the probability of observing an event time T larger than some given t. A basic property of the survival function is that $S(t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function. Because of this relationship, $S(t)$ is assumed to go to zero as time proceeds, which means that all subjects under observation are expected to experience the event of interest eventually. As opposed to medical science where the event of interest is usually death, this property does not seem valid in the credit risk context, as a substantial part of the population will never experience default. In fact, it can be argued that insusceptibility to default is the main

reason behind the high censoring rate. The proportion of observations where default is not observed might in practice even exceed 95%.

A remedy for this, the so-called "mixture cure model", was initially proposed by Farewell (1982) to model long-term survivors in the medical context. This model contains a logistic regression component, modeling "insusceptibility" to the event of interest, and a survival component, modeling the survival times of an individual conditioning on susceptibility. While using parametric survival distributions in the survival component initially, Kuk and Chen (1992) extended the mixture cure model using non-parametric survival distributions (see also Peng and Dear, 2000; Sy and Taylor, 2000a). Cai et al. (2012b) introduced the `smcure`-package in R (R Core Team, 2014) to estimate semi-parametric mixture cure models. This latter version of the mixture cure model was introduced in the credit risk context for the first time by Tong et al. (2012). Dirick et al. (2015) developed a model selection criterion for these models, and applied this to credit risk data.

While the use of TVCs has been investigated in (non-mixture) survival models, both in medical research (see among others Andersen, 1992) and in the credit context (see Bellotti and Crook, 2009), to our knowledge TVCs have not been implemented before in mixture cure models. In the present chapter, we therefore examine TVCs in these models, more specifically macro-economic factors, along with the usual time-independent covariates.

The remainder of this chapter is organized as follows. In Section 4.2, we give a short overview on which types of different TVCs exist. In Sections 4.3, 4.4 and 4.5, we discuss the mixture cure model with TVCs, the likelihood function and computational details. The simulation setup and results are discussed in Section 4.6, and a credit risk data example is presented in Section 4.7. Section 4.8 concludes.

4.2 Time-varying covariates

4.2.1 Internal versus external TVCs

TVCs can be segmented into two general classes: internal and external TVCs, see, among others, Kalbfleisch and Prentice (2002, Chapter 6), Hosmer et al. (2008, Chapter 7) and Cortese and Andersen (2010). An internal TVC is one whose value is typically subject-specific and requires the subject to be under direct observation. An example of a TVC in the credit risk context is a customer's current account balance, or a patient's cholesterol level in the medical context. From the bio-medical point of view, an internal covariate generally requires the survival of the individual for its existence. In this sense, the internal TVC-path carries direct information on the timing of the event if this event is death.

An external TVC on the other hand does not require subjects to be under direct observation, nor does its existence depend on the occurrence of the event of interest. Examples of external TVCs are the inflation rate in the credit risk context, and air pollution in the biomedical context. In general, these TVC-types are usually environmental factors that apply to all subjects under observation; however, subject-specific properties such as age are considered to be external as, given a subject's birth date, age can be determined at any time. A time-fixed covariate can be seen as a special case of an external time-dependent covariate, where its value is measured in advance and fixed for the entire study, e.g. the applicant's bureau score.

Formally, in a non-mixture survival context, denote $\boldsymbol{x}_i(t) = (x_{i1}(t), \ldots, x_{il}(t))$ the covariate vector at time t for individual $i = 1, \ldots, n$. Additionally, denote the covariate history up to time t: $X_i(t) = \{\boldsymbol{x}_i(u);\ 0 \leq u < t\}$. The available information for each observation i is given by the time $T_i = \min(U_i, C_i)$, where U_i denotes the true event time and C_i is the censoring time, a corresponding censoring indicator $\delta_i = I(U_i \leq C_i)$ and $X_i(t_i)$, the covariate history until t_i.

A TVC is external when for all v, t, such that $0 < v \leq t$ it satisfies the

condition (Kalbfleisch and Prentice, 2002, Chapter 6)

$$P\left(T \in [v, v+dv) \,|\, X(v),\ T \geq v\right) = P\left(T \in [v, v+dv) \,|\, X(t),\ T \geq v\right). \quad (4.1)$$

The rationale behind this condition is that although a time-dependent covariate may influence the event rate over time, its future path until any time t is not affected by the occurrence of the event of interest at time v. The difference between internal and external covariates has great implications on survival function estimation. In the presence of external covariates, the standard relationship between the survival function and the hazard function,

$$S\left(t \mid X(t)\right) = P\left(T \geq t \mid X(t)\right) = \exp\left(-\int_0^t \lambda\{v \mid X(v)\} dv\right) \quad (4.2)$$

holds. All TVCs described in this chapter are macro-economic variables and are all external, hence (4.2) can be used.

However, bear in mind that in case of internal covariates, extra attention should be given to the estimation of the survival function, see Andersen (1992) for a probabilistic model for survival function estimation in presence of internal TVCs. Because of the nature of internal TVC's and non-compliance to (4.1), however, estimation of instantaneous hazards is possible, but cumulative hazards and survival probabilities are no longer feasible through (4.2). To see this, we reconsider the example of the internal covariate cholesterol level in a study where the event of interest is death. Looking at (4.2), any measurable cholesterol level value would indicate that the subject under investigation is still alive, hence, $S\left(t \mid X(t)\right) = P\left(T \geq t \mid X(t)\right) = 1$ given that $X(t)$ is measurable. For more information on this issue, we refer to Kalbfleisch and Prentice (2002, Chapter 6) and Fisher and Lin (1999).

4.2.2 Macro-economic factors

Being a function of a (continuous) time t, TVCs can theoretically change continuously. This is approximately the case for some macro-economic

variables (e.g. stock prices), others tend to be documented over longer periods of times such as unemployment rates (weekly, monthly or yearly). To manage TVCs in survival models, the observation period of each subject is split in several time-periods, which are defined by adjacent event times (Fox, 2002). Let $x_{ip}(t)$ (where $p \in \{1, \ldots, l\}$) be one specific time-dependent covariate, and let $B_1 < \ldots < B_m$ be all the unique event or censoring times observed in the data set. To manage the data, subject i must have exactly one TVC value for each of the intervals $\{(0, B_1], (B_1, B_2], \ldots, (B_{k_i-1}, B_{k_i}]\}$, where $k_i \in \{1, \ldots, m\}$ and $B_{k_i} = t_i$, hence each subject has its own set of TVC values until its own censoring or event time t_i.

Applied to default events in loans, these specific intervals represent the respective number of months a subject has been repaying until default or censoring. As a result, the TVCs under investigation are the monthly averages of specific macro-economic factors. This is denoted by replacing $x_{ip}(t)$ by $\bar{x}_{ip}(t) = \left(\bar{x}_{ip}((0, B_1]), \bar{x}_{ip}((B_1, B_2]), \ldots, \bar{x}_{ip}((B_{k_i-1}, B_{k_i}])\right)$, where $\bar{x}_{ip}((B_{j-1}, B_j])$ is the average value of TVC p for subject i over the time interval $(B_{j-1}, B_j]$.

4.3 A mixture cure model with TVCs

In a mixture cure model, cases are categorized into two groups: a group that will experience the event, and a group of so-called 'insusceptible' cases that will not experience the event of interest. These groups are modeled using a mixture distribution where a logistic regression model provides a mixing proportion of the insusceptible cases and where a survival model describes the cases susceptible to the event of interest (Tong et al., 2012). In the credit risk context, where the event of interest is loan default, every event-type that is not default (e.g. loan maturity, early repayment) is considered as censored. By consequence, there is heavy right-censoring and a large group of insusceptible cases is expected to be present.

For each subject i, the censoring indicator δ_i denotes whether subject i experiences the event of interest during the observation period ($\delta_i = 1$), or not ($\delta_i = 0$). This censoring indicator provides partial information on sus-

ceptibility; however, when an observation is censored, it is unclear whether the event will still occur after the observation period has terminated. Introducing susceptibility indicator Y_i, where $Y_i = 1$ when an observation is susceptible and $Y_i = 0$ if not, three different combinations of Y_i and δ_i are possible:

1) $Y_i = 1$ and $\delta_i = 1$: uncensored and susceptible, the event takes place during observation period;
2) $Y_i = 1$ and $\delta_i = 0$: censored and susceptible, the event will take place, however is not observed;
3) $Y_i = 0$ and $\delta_i = 0$: censored and insusceptible, the event is not observed and will never take place.

For each observation i, T_i and δ_i are fully observed, Y_i is only observed and equal to 1 when $\delta_i = 1$.

4.3.1 The model

In a model with both time-dependent covariates $\boldsymbol{x}(t)$ and time-fixed covariates $\boldsymbol{z}$, the unconditional survival function of the mixture cure model, for given values of the covariates $\boldsymbol{x}(t)$ and $\boldsymbol{z}$, is given by

$$S(t \mid \boldsymbol{z}, \boldsymbol{x}(t)) = \pi(\boldsymbol{z}) S(t \mid Y = 1; \boldsymbol{z}, \boldsymbol{x}(t)) + 1 - \pi(\boldsymbol{z}). \tag{4.3}$$

The so-called 'incidence model', $\pi(\boldsymbol{z}) = P(Y = 1; \boldsymbol{z})$, is the proportion of susceptible accounts given covariate vector $\boldsymbol{z} = (z_1, \ldots, z_s)$, modeled using a binary logit,

$$\pi(\boldsymbol{z}) = \frac{\exp(\boldsymbol{b'z})}{1 + \exp(\boldsymbol{b'z})}. \tag{4.4}$$

Note that, in this part of the mixture cure model, only time-fixed covariates are incorporated. The conditional survival function is modeled using a semi-parametric proportional hazard regression model such that

$$\begin{aligned} & S(t \mid Y = 1; \boldsymbol{z}, \boldsymbol{x}(t)) \\ &= \exp\Big(- \exp(\boldsymbol{\beta'z} + \boldsymbol{\beta}_t' \boldsymbol{x}(t)) \int_0^t h_0(u \mid Y = 1) du \Big), \end{aligned} \tag{4.5}$$

with h_0 the unspecified baseline hazard function, $\boldsymbol{x}(t) = (x_1(t), \ldots, x_l(t))$ a l-vector of time-dependent covariates and $\boldsymbol{z} = (z_1, \ldots, z_s)$ a time-fixed covariate vector identical to the one in the incidence model. Note that from a theoretical point of view, the incidence and latency time-fixed covariate vectors may contain different variables; however in this chapter, focusing on time-dependent covariates, these covariates are kept equal in all practical examples. For mixture cure models with different time-fixed covariate elements in latency and incidence models, we refer to Dirick et al. (2015).

4.3.2 The likelihood function

In order to construct the likelihood function, more attention should be devoted to the TVCs. When TVCs are present in the data set, the biggest challenge lies in data management in order to make TVC-handling possible. In practice, this is done by requiring that each time period (bounded by $B_1 < \ldots < B_m$, see section 4.2.2) for a specific individual appears in a separate row in the data set (Fox, 2002). Denote $\lambda_{i,j}$ the interval-specific censoring indicator for interval $j \in \{1, \ldots, k_i\}$ of observation i. The complete likelihood, given full information on Y, can be expressed as:

$$\begin{aligned} L_c(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\beta}_t) &= \prod_{i=1}^{n} \big(1 - \pi(\boldsymbol{z}_i)\big)^{(1-Y_i)} \pi(\boldsymbol{z}_i)^{Y_i} \\ &\quad \times \prod_{j=1}^{k_i} h(t_j \,|\, Y_i = 1; \boldsymbol{z}_i, \boldsymbol{x}_i(t_j))^{\lambda_{i,j} Y_i} S(t_j \,|\, Y_i = 1; \boldsymbol{z}_i, \boldsymbol{x}_i(t_j))^{Y_i} \end{aligned}$$

where $h(t_j \mid \cdot)$ and $S(t_j \mid \cdot)$ are, respectively, the hazard and survival contributions at the time point given by the upper bound B_j of the corresponding interval , and $\boldsymbol{x}_i(t_j)$ is the value of the TVC of observation i in the interval $(B_{j-1}, B_j]$. The log likelihood function can then be written as the sum of the latency and incidence log likelihoods,

$$\begin{aligned} &\log L_c(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\beta}_t \mid Y; \boldsymbol{z}, \boldsymbol{x}(t)) \\ &\quad = \log L_{inc}(\boldsymbol{b} \mid Y; \boldsymbol{z}) + \log L_{lat}(\boldsymbol{\beta}, \boldsymbol{\beta}_t \mid Y; \boldsymbol{z}, \boldsymbol{x}(t)), \end{aligned} \tag{4.6}$$

where

$$\log L_{inc}(\boldsymbol{b}\,|\,Y;\boldsymbol{z}) \;=\; \sum_{i=1}^{n}(1-Y_i)\big(1-\pi(\boldsymbol{z}_i)\big)+Y_i\,\pi(\boldsymbol{z}_i) \qquad (4.7)$$

$$\begin{aligned}\log L_{lat}(\boldsymbol{\beta},\boldsymbol{\beta}_t\,|\,Y;\boldsymbol{z},\boldsymbol{x}(t)) \;=\; &\sum_{i=1}^{n}\sum_{j=1}^{k_i}Y_i\lambda_{i,j}\log h(t_j\,|\,Y_i=1;\boldsymbol{z}_i,\boldsymbol{x}_i(t_j))\\ &+Y_i\log S(t_j\,|\,Y_i=1;\boldsymbol{z}_i,\boldsymbol{x}_i(t_j)).\end{aligned} \qquad (4.8)$$

As noted at the start of Section 4.3, Y_i is missing for the censored cases. As we do not have an exact expression for $\log L_c(\boldsymbol{b},\boldsymbol{\beta},\boldsymbol{\beta}_t \mid Y;\boldsymbol{z},\boldsymbol{x}(t))$, the expectation maximization-algorithm (EM algorithm) will be used. The EM algorithm is an iterative procedure to find the maximum likelihood-estimates of an underlying distribution from data that are incomplete (Dempster et al., 1977). We provide the needed adjustments to the algorithm to incorporate TVCs in mixture cure models.

4.4 Implementation using the EM-algorithm

4.4.1 The E-step

Denote the parameter-triplet $(\boldsymbol{b},\boldsymbol{\beta},\boldsymbol{\beta}_t)$ by Θ, and the observed information for each observation $(\lambda_{i,j},\delta_i,t_i)$ by $\mathbf{O}$. The conditional expectation of the complete-data log likelihood (formula (4.6)) in the $(r+1)$th E-step is given by

$$Q(\Theta^{(r+1)} \mid \Theta^{(r)}) = E[\log L_c(\boldsymbol{\Theta}^{(r+1)} \mid Y;\boldsymbol{z},\boldsymbol{x}(t)) \mid \mathbf{O},\boldsymbol{\Theta}^{(r)}]. \qquad (4.9)$$

It can easily be seen that these functions are linear in Y_i, which reduces the problem to find an expression for the conditional expectation of Y_i, which is given by

$$\begin{aligned} w_i^{(r)} \;&=\; E(Y_i \mid \mathbf{O},\boldsymbol{\Theta}^{(r)})\\ &=\; \begin{cases}\dfrac{\pi(\boldsymbol{z}_i)S(t_i \mid Y_i=1;\boldsymbol{z}_i,\boldsymbol{x}_i(t_i))}{\pi(\boldsymbol{z}_i)S(t_i \mid Y_i=1;\boldsymbol{z}_i,\boldsymbol{x}_i(t_i))+(1-\pi(\boldsymbol{z}_i))} & \text{for } \delta_i=0\\ 1 & \text{for } \delta_i=1.\end{cases}\end{aligned} \qquad (4.10)$$

Note that $E(Y_i \mid \mathbf{O}, \boldsymbol{\Theta}^{(r)})$ takes one value per iteration for each observation. The weights $w_i^{(r)}$ are computed using the value of the TVCs at time of censoring, and can be interpreted as the probability that individual i will be susceptible to the event.

4.4.2 The M-step

In the M-step, the expected complete-data log likelihood in (4.9) is maximized with respect to the unknown parameters. The conditional expectation of the incidence log likelihood is straightforward, replacing Y_i's in (4.7) by $w_i^{(r)}$. The conditional expectation of the latency log likelihood (4.8) can then be written, using $\lambda_{i,j} \log w_i^{(r)} = 0$ and $\lambda_{i,j} w_i^{(r)} = \lambda_{i,j}$ (Cai et al., 2012a) as

$$
\begin{aligned}
&E[\log L_{lat}(\boldsymbol{\beta}, \boldsymbol{\beta}_t \mid Y; \boldsymbol{z}, \boldsymbol{x}(t)) \mid \mathbf{O}, \boldsymbol{\Theta}^{(r)}] \\
&\quad = \sum_{i=1}^{n} \sum_{j=1}^{k_i} \lambda_{i,j} \log \left(w_i^{(r)} h(t_j \mid Y_i = 1; \boldsymbol{z}_i, \boldsymbol{x}_i(t_j)) \right) \\
&\qquad\qquad + w_i^{(r)} \log S(t_j \mid Y_i = 1; \boldsymbol{z}_i, \boldsymbol{x}_i(t_j)) \\
&\quad = \log \prod_{i=1}^{n} \prod_{j=1}^{k_i} \left(w_i^{(r)} h_0(t_j) \exp(\boldsymbol{\beta}' \boldsymbol{z}_i + \boldsymbol{\beta}_t' \boldsymbol{x}_i(t_j)) \right)^{\lambda_{i,j}} \\
&\qquad\qquad \times \left(S_0(t_j)^{\exp(\boldsymbol{\beta}' \boldsymbol{z}_i + \boldsymbol{\beta}_t' \boldsymbol{x}_i(t_j))} \right)^{w_i^{(r)}} \\
&\quad = \log \prod_{i=1}^{n} \prod_{j=1}^{k_i} \left(h_0(t_j) \exp(\boldsymbol{\beta}' \boldsymbol{z}_i + \boldsymbol{\beta}_t' \boldsymbol{x}_i(t_j) + \log w_i^{(r)}) \right)^{\lambda_{i,j}} \\
&\qquad\qquad \times \left(S_0(t_j)^{\exp(\boldsymbol{\beta}' \boldsymbol{z}_i + \boldsymbol{\beta}_t' \boldsymbol{x}_i(t_j) + \log w_i^{(r)})} \right).
\end{aligned}
$$

When (4.9) is maximized, the baseline survival function of the r^{th} M-step should be updated in order to proceed with the next E-step. This is done non-parametrically using the Breslow-type estimator for $S_0(t)$ and combining the results of Andersen (1992) and Cai et al. (2012a). Denote $R(t_j)$ the individuals at risk in the interval $(B_{j-1}, B_j]$, then

$$
\hat{S}_0(t) = \exp\left(- \sum_{j: t_j \leq t} \frac{\sum_{i \epsilon R(t_j)}^{n} \lambda_{i,j}}{\sum_{i \epsilon R(t_j)}^{n} w_i^{(r)} \exp(\boldsymbol{\beta}'^{(r)} \boldsymbol{z}_i + \boldsymbol{\beta}_t'^{(r)} \boldsymbol{x}_i(t_j))} \right). \quad (4.11)
$$

The E-step and the M-step are repeated until parameter convergence.

4.4.3 Variance estimation

When estimating parameters through the EM-algorithm, standard errors of these parameter estimates are not directly available. A widespread method for estimating variances in the mixture cure context is bootstrap (e.g. Peng, 2003; Cai et al., 2012a; Tong et al., 2012). Though the bootstrap technique is easy to implement, this method is computationally extensive, especially with big datasets and resulting slow convergence of the EM-algorithm. In this work we use the supplemented EM (SEM)-algorithm introduced by Meng and Rubin (1991), as in Section 2.3.3. While other approximation methods exist, see, among others Sy and Taylor (2000a); Peng and Dear (2000), the advantage of SEM is that it can be applied to any problem to which EM is applied, assuming that there is access to the complete-data asymptotic variance-covariance matrix, which is indeed the case here.

4.5 Computational scheme

4.5.1 Data structure

Including TVCs in the survival part of the mixture cure model requires rearrangement of the data. To make TVCs computationally feasible in a Cox PH model, each time period $(B_{j-1}, B_j]$ with $j = 1, \ldots, k_i$ for each individual i is represented as a single row in the data set (Fox, 2002). Note that the number of rows for each observation depends on the observation itself as $B_{k_i} = t_i$. The advantage of this data structure is that one can use the `coxph`-function in package `survival` in R (Therneau, 2014), using preamble "`Surv(start, stop, default)`" instead of the more familiar "`Surv(time, default)`".

The mixing proportions of the mixture cure model modeled by the binomial logit do not include TVCs, and using several lines per observation in this model part would lead to wrong $\boldsymbol{b}$-parameter estimates. As a result, for the mixture cure model with TVCs, different data set structures are

	δ_i	t_i	$\boldsymbol{z}_1$	$\boldsymbol{z}_2$
obs 1	1	1	-1	2
obs 2	0	2	0.3	3
obs 3	1	3	0.4	2.3

	B_{j-1}	B_j	$\lambda_{i,j}$	δ_i	t_i	$\boldsymbol{z}_1$	$\boldsymbol{z}_2$	$\boldsymbol{x}_1(t)$	$\boldsymbol{x}_2(t)$
obs 1	0	1	1	1	1	-1	2	0.3	-0.7
obs 2	0	1	0	0	2	0.3	3	0.2	0.4
obs 2	1	2	0	0	2	0.3	3	0.7	-0.1
obs 3	0	1	0	1	3	0.4	2.3	0.5	-1
obs 3	1	2	0	1	3	0.4	2.3	0.2	-0.3
obs 3	2	3	1	1	3	0.4	2.3	0.4	0.2

Table 4.1: *Example of the incidence versus latency model data structure. The left hand table represents the data structure for the binomial logit part of the mixture cure model, where no TVCs are present. The right hand table is the long data structure needed for incorporation of TVCs in the survival part of the model.*

being used depending on whether the respective calculations are performed on the latency versus the incidence part of the model. An example of the 'short' (incidence) data structure versus the 'long' (latency) data structure is represented in Table 4.1. To transform the general short form of survival data into the long structure, Fox and Carvalho (2012) introduced the "unfold" function in the R-package `RcmdrPlugin.survival`.

4.5.2 Procedure

The procedure consists of three main steps: the initialization step, the E-step and the M-step.

Initialization

1) *Initialize w:* Initialize $w_i^{(0)}$ by taking $w_i^{(0)} = \delta_i$. Each observation has one $w_i^{(0)}$.

2) *Initialize $\boldsymbol{b}$:* Fit a binomial logit model to $w_i^{(0)}$ using the 'short' data set and covariate vector $\boldsymbol{z}$, in order to retrieve an initial estimate $\hat{\boldsymbol{b}}^{(0)}$.

3) *Initialize $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_t$:* Obtain $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{\beta}}_t^{(0)}$-estimates using the `coxph`-function for the (long) survival data including TVCs. Use w_i's as

weights in the model, matching w_i with each line that corresponds with observation i.

4) *Initialize $S_0(t)$:* Compute $\hat{S}_0^{(0)}(t)$ using formula (4.11).

Expectation step

1) Compute $\pi_i^{(1)}(z_i)$ for each i, using Formula (4.4) and $\hat{\boldsymbol{b}}^{(0)}$.

2) Compute $w_i^{(1)}$ for each i, using Formula (4.10) $\hat{\beta}^{(0)}$. Note that the survival estimates used here,

$$\hat{S}^{(0)}(t_i \mid Y_i = 1; \boldsymbol{z}_i, \boldsymbol{x}_i(t_i)) = \hat{S}_0^{(0)}(t_i)^{\exp(\hat{\boldsymbol{\beta}}'^{(0)}\boldsymbol{z}_i + \hat{\boldsymbol{\beta}}_t'^{(0)}\boldsymbol{x}_i(t_i))},$$

correspond for each observation with the estimate at time of the last observation, hence the linear predictor consists of the TVC-values at time t_i.

Maximization step

1) *Update $\boldsymbol{b}$:* Obtain a new estimate $\hat{\boldsymbol{b}}^{(1)}$ using the $w_i^{(1)}$ of the E-step when fitting the binomial logit model.

2) *Update $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_t$:* Obtain a new estimate of $\hat{\boldsymbol{\beta}}_t^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(1)}$ including the $w_i^{(1)}$'s as weights.

3) *Update S*: Obtain a new estimate of $\hat{S}^{(0)}(t)$ using formula (4.11).

The E- and M-step will be repeated with all updated estimates, until parameter convergence. The algorithm stops when the sum of the squared differences between $\left(\hat{\boldsymbol{\beta}}_t^{(r+1)}, \hat{\boldsymbol{\beta}}_t^{(r+1)}, \hat{\boldsymbol{b}}^{(r+1)}\right)$ and $\left(\hat{\boldsymbol{\beta}}_t^{(r)}, \hat{\boldsymbol{\beta}}_t^{(r)}, \hat{\boldsymbol{b}}^{(r)}\right)$ is smaller than a pre-specified value.

4.6 Simulation study

4.6.1 Simulating survival times with time-dependent covariates

In our simulation study, we include both time-fixed covariates $\boldsymbol{z}$ (associated with $\boldsymbol{b}$ and $\boldsymbol{\beta}$) and TVCs $\boldsymbol{x}(t)$ (associated with $\boldsymbol{\beta}_t$). If one simulates survival times using the exponential distribution with time-invariant covariates only, the survival times and the cumulative hazard function can be defined piecewise,

$$T = -\frac{\log(u)}{\lambda \exp(\boldsymbol{\beta}'\boldsymbol{z})}, \quad H(-\log(u), z) = \lambda \exp(\boldsymbol{\beta}'\boldsymbol{z})(-\log(u))$$

where $u \sim U(0, 1)$.

Austin (2012) describes a method for generating survival times in the presence of TVCs. In this work, TVCs are constrained to be dichotomous variables with a limited number of changes between 0 and 1. For our purpose, we generalized this setting in two ways:

1) The TVC can change value from one time period to another, where a time period is defined by two adjacent event or censoring times.

2) The TVC can take any value, and does not need to be dichotomous.

When running a simulation, first, the boundaries that define the TVC-intervals should be chosen. We denote B_j as the timepoints where the covariate values change. Note that $j \in (1, \ldots, m)$ with $m \leq n - 1$, with n the number of cases, as both event and censoring times are unique in a simulation study when using continuous time distributions.

As a notational convention, we use $x(t_j)$ for the value of the time-dependent covariate in the interval $(B_{j-1}, B_j]$. In a generalization of the simulation method by Austin (2012), the cumulative hazard function is given by, denote $h = -\log(u)$,

$$
\begin{aligned}
&H(h,z,x(t))\\
&= \begin{cases}
\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\beta}_t'\boldsymbol{x}(t_1))(h) & \text{if } h\le B_1\\
\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z})\Big[\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_1))B_1+\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_2))(h-B_1)\Big] & \text{if } B_1<h\le B_2\\
\vdots & \\
\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z})\Big[\sum_{j=1}^{m}\big(\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_j))(B_j-B_{j-1})\big)+\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_{m+1}))(h-B_m)\Big] & \text{if } B_m<h
\end{cases}
\end{aligned}
$$

The domain of the cumulative hazard function can be divided into mutually exclusive intervals $D_1=(0,B_1], D_2=(B_1,B_2],\ldots,D_{m+1}=(B_m,\infty)$. The range of the corresponding cumulative hazard functions is

$$
\begin{aligned}
R_1 &= \Big(0,\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\beta}_t'\boldsymbol{x}(t_1))B_1\Big];\\
R_2 &= \Big(\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\beta}_t'\boldsymbol{x}(t_1))B_1,\\
&\qquad \lambda\exp(\boldsymbol{\beta}'\boldsymbol{z})\{\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_1))B_1+\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_2))(B_2-B_1)\}\Big];\\
&\vdots\\
R_{m+1} &= \Big(\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z})\sum_{j=1}^{m}\big(\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_j))(B_j-B_{j-1})\big),\infty\Big).
\end{aligned}
$$

Inverting each of the piecewise components of the cumulative hazard function we can simulate the survival time as $H^{-1}(h,z,x)$ with

$$
\begin{aligned}
&H^{-1}(h,z,x(t))\\
&= \begin{cases}
\dfrac{h}{\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\beta}_t'\boldsymbol{x}(t_1))} & \text{if } h\in R_1\\
\dfrac{h-\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\beta}_t'\boldsymbol{x}(t_1))B_1+\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\beta}_t'\boldsymbol{x}(t_2))B_1}{\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\beta}_t'\boldsymbol{x}(t_2))} & \text{if } h\in R_2\\
\vdots & \\
\dfrac{h+\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z})\{\sum_{j=1}^{m}(-\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_j))(B_j-B_{j-1}))+\exp(\boldsymbol{\beta}_t'\boldsymbol{x}(t_{m+1}))(B_m)\}}{\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\beta}_t'\boldsymbol{x}(t_{m+1}))} & \text{if } h\in R_{m+1}.
\end{cases}
\end{aligned}
$$

4.6.2 Simulation setup and results

Uncorrelated time-varying covariates

In the simulation study, the probability of being insusceptible is generated using a logistic model where $\pi(z)=\dfrac{\exp(\boldsymbol{b}'\boldsymbol{z})}{1+\exp(\boldsymbol{b}'\boldsymbol{z})}$, and the survival times

	true value	Mean est value	Mean std	Bias	MSE
$\hat{b}_0$	2	2.0758	0.2539	0.0758	0.0937
$\hat{b}_1$	0.5	0.5400	0.1530	0.0400	0.0270
$\hat{b}_2$	-2.3	-2.3678	0.1482	0.0678	0.0789
$\hat{\beta}_1$	-1.2	-1.1504	0.0697	0.0496	0.0086
$\hat{\beta}_2$	1	0.8941	0.0769	0.1059	0.0220
$\hat{\beta}_{t1}$	1	0.9882	0.0716	0.0118	0.0055
$\hat{\beta}_{t2}$	-0.7	-0.6915	0.0714	0.0085	0.0056

Table 4.2: *Results for simulation setting 1. Insusceptible= 22.72%, Censoring= 32.85%.*

of susceptible cases are generated using an exponential distribution with $\lambda = 0.7$. Two uncorrelated time-fixed covariates $z_1 \sim N(1.5, 0.6)$ and $z_2 \sim \text{bin}(1, 0.5)$, and two time-dependent covariates $x_1(t) \sim N(2, 0.5)$ and $x_2(t) \sim N(0.8, 0.5)$ are generated.

We used two settings, Setting 1 with low insusceptibility and Setting 2 with high insusceptibility, and corresponding low and high censoring. Susceptibility is managed through the generating b-parameters, (2, 0.5, -2.3) and (-1.5, 0.5, -2) respectively. Censoring times are generated from an exponential distribution as well, however using $\lambda = 0.1$ and $\lambda = 0.2$ respectively. For each of the two settings, results are based on n=1000 and with 100 replications. Note that, as stated in section 4.6.1, the TVC can theoretically change value $n-1$ times. To imitate real-data situations, we constrained the TVCs to change values 60 times at most, as the data sets we typically use have a loan term of five years or less.

In Table 4.2 and 4.3, the true generating parameter values are shown, as well as the mean parameter estimates and standard errors over the 100 simulation runs. Additionally, the absolute biases and the mean squared errors between parameter estimates and true values are given. Comparing the results of Table 4.2 with Table 4.3, it can be seen that higher censorship, generally leads to higher bias and MSE. However, biases remain low in setting 2, and especially the parameters related to the TVCs, β_{t1} and β_{t2}

	true value	Mean est value	Mean std	Bias	MSE
$\hat{b}_0$	-1.5	-1.3790	0.2128	0.1210	0.1262
$\hat{b}_1$	0.5	0.5210	0.0925	0.0210	0.0610
$\hat{b}_2$	-2	-2.0909	0.1868	0.0909	0.0716
$\hat{\beta}_1$	-1.2	-1.0636	0.1483	0.1364	0.0686
$\hat{\beta}_2$	1	0.8658	0.2042	0.1342	0.0960
$\hat{\beta}_{t1}$	1	0.9427	0.1586	0.0573	0.0249
$\hat{\beta}_{t2}$	-0.7	-0.6858	0.1585	0.0142	0.0271

Table 4.3: *Results for simulation setting 2. Insusceptible= 80.71%, Censoring= 86.1%.*

are well-estimated. This was a general result while running simulations. The abundant information in the TVCs (for each case in our simulation runs 40 to 60 different values for one TVC) enables an accurate parameter calculation.

Correlated time-varying covariates

In real life, macro-economic factors are all linked and influence each other. To mimic this behaviour, a situation is tested where TVCs are highly correlated in simulation setting 3. Time-fixed covariates have the same distributions as for setting 1 and 2, and generating parameters are as in setting 1. However, this time we included three TVCs that are highly correlated, with mean and covariance matrix

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix}; \qquad \mu = \begin{pmatrix} 2 \\ 0.8 \\ -0.7 \end{pmatrix}; \qquad \Sigma = \begin{pmatrix} 0.7 & 0.8 & 0.8 \\ 0.8 & 1.2 & 0.8 \\ 0.8 & 0.8 & 1.0 \end{pmatrix}.$$

The corresponding correlation matrix for the time-dependent covariates is then given by

$$\rho = \begin{pmatrix} 1 & 0.873 & 0.956 \\ 0.873 & 1 & 0.730 \\ 0.956 & 0.730 & 1 \end{pmatrix}.$$

	true value	Mean est value	Mean std	Bias	MSE
$\hat{b}_0$	2	2.0834	0.2654	0.0834	0.1273
$\hat{b}_1$	0.5	0.6053	0.1562	0.1053	0.0592
$\hat{b}_2$	-2.3	-2.3036	0.1682	0.0036	0.1363
$\hat{\beta}_1$	-1.2	-1.0529	0.0728	0.1471	0.0303
$\hat{\beta}_2$	1	0.6866	0.0769	0.3134	0.1189
$\hat{\beta}_{t1}$	1	0.9276	0.2716	0.0724	0.1048
$\hat{\beta}_{t2}$	-0.7	-0.6663	0.0924	0.0337	0.0120
$\hat{\beta}_{t3}$	0.5	0.5036	0.1673	0.0036	0.0367

Table 4.4: *Results for simulation setting 3. Insusceptible= 22.72%, Censoring= 39.01%*

In setting 3, censoring times are generated using an exponential distribution with $\lambda = 0.15$. The results in Table 4.4 show that, though mean standard errors of the TVCs are higher in comparison with the time-fixed covariates and the results in Tables 4.2 and 4.3, biases and mean squared errors are not notably higher for the b-parameters and the β_t-parameters. β_1 and β_2, however, do seem to have a higher bias. This result is due to the fact that this setting resulted in a higher gap between censored percentage and insusceptible percentage ($39.01 - 22.72\% = 16.29\%$). Throughout our simulations, it has become clear that the β-parameter estimates in generally deteriorate when this gap becomes lot bigger than 10%. The β_t-parameter estimates, however, do not seem to be affected.

4.7 Data set with macro-economic variables

The data used was provided by a major Belgian financial institution. The sample, consisting of 20 000 loans with a fixed loan term of 36 months, spanned a period of loans that were initiated between January 2004 and May 2014. In each of the models that are discussed, seven time-independent covariates (described in Table 4.5) are included as a baseline. Additionally, six macro-economic factors were gathered through the online database

	Description	Type
z_1	Annual income (per 1000)	continuous
z_2	Age	continuous
z_3	Monthly child allowance (Y/N)	categorical
z_4	Number of years at current address	continuous
z_5	Total employment years	continuous
z_6	Bureau score	continuous
z_7	Mortgage on real estate (Y/N)	categorical

Table 4.5: *Credit loan data, description of the time-independent covariates.* $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_4$ *and* $\boldsymbol{z}_5$ *are mean-centered,* $\boldsymbol{z}_6$ *log transformed.*

from the Belgian National Bank (NBB, 2015). A TVC-value was retained for each month in the years 2004 until 2014, correcting for both trend and seasonality by taking the yearly difference for each TVC (e.g. the TVC-value for unemployment in August 2008 is the difference between its value in August 2008 and August 2007). As some macro-economic factors may have a delayed effect on default, timelags of six months were introduced for the TVCs of market interest rate, inflation rate and unemployment. Hence, we examine the effect of the inflation rate in, say, February 2005 on possible default in August 2005.

4.7.1 Data analysis using the mixture cure model

Information about the time-independent and time-dependent covariates can be found in Tables 4.5 and 4.6 respectively. Several mixture cure models, each including the same seven time-independent covariates, and three or four different TVCs (leading to thirty-five models in total) were analyzed. Including a couple of TVCs each time was preferred over including all at once, to ensure again that correlation issues can not play a role. Having a resulting range of models, it is easy to compare covariate values of different macro-economic factors over the different models as a stability test. The resulting parameter estimates for the TVCs can be found in Ta-

	Type	Lag	Description
$\bar{x}_1(t)$	Interest Rate	6 months	As the interest rates of the Belgian financial institution were not disclosed, the minimum bid interest rate was chosen. This refers to the minimum interest rate at which counterparties may place their bids for refinancing operations.
$\bar{x}_2(t)$	BEL 20 index	none	The benchmark stock market index of Euronext Brussels, consisting of ten to twenty (depending on the period) companies that are traded at Brussels Stock Exchange. The TVC is expressed as the difference between the index of the current period and the previous year, divided by 1000.
$\bar{x}_3(t)$	Consumer confidence	none	Monthly survey on a variable sample of 1850 households conducted by the National Bank. The survey, harmonised at European level, supplies information on the appreciation of the consumers regarding the progress of the economy in general and regarding their own situation in particular.
$\bar{x}_4(t)$	GDP	none	Growth in the Belgian Gross Domestic Product with respect to the same period in the previous year (GDP growth is documented quarterly).
$\bar{x}_5(t)$	Inflation rate	6 months	Percentage changes in consumer price compared to the corresponding period of the previous year.
$\bar{x}_6(t)$	Unemployment	6 months	Harmonised data derived from the Labour Force Survey (LFS, population older than 15 years), monthly adjusted by using the administrative national unemployment figures, in accordance with the Eurostat methodology.

Table 4.6: *Time-dependent covariates $\bar{x}_1(t)-\bar{x}_6(t)$ are differential macro-economic factors that change month by month. A specific TVC is the difference between the nominal macro-economic factor value in a specific month and the same factor twelve months before.*

	interest rate	BEL 20 index	cons confidence	GDP	inflation rate	unemployment	AIC_{cd}
model 1	0.056 (0.036)	-0.107 (0.052)$^{(*)}$	0.002 (0.005)				29706.65
model 2	0.067 (0.033)$^{(*)}$	-0.133 (0.058)$^{(*)}$		0.02 (0.015)			29858.27
model 3	0.073 (0.037)$^{(\cdot)}$	-0.134 (0.055)$^{(*)}$			-0.044 (0.037)		29677.70
model 4	0.055 (0.042)	-0.092 (0.043)$^{(*)}$				0.007 (0.06)	29706.46
model 5	0.049 (0.036)		-0.003 (0.005)	-0.004 (0.017)			29412.13
model 6	0.053 (0.037)		-0.005 (0.005)		-0.007 (0.034)		29413.14
model 7	0.043 (0.042)		-0.004 (0.004)			-0.021 (0.056)	29417.41
model 8	0.05 (0.02)$^{(*)}$			-0.014 (0.028)	-0.005 (0.04)		29413.69
model 9	0.036 (0.043)			-0.011 (0.013)		-0.032 (0.054)	29450.2
model 10	0.039 (0.043)				0.012 (0.027)	-0.035 (0.055)	29441.97
model 11		-0.11 (0.064)$^{(\cdot)}$	-0.001 (0.005)	0.006 (0.018)			30163.56
model 12		-0.109 (0.061)$^{(\cdot)}$	-0.001 (0.005)		-0.014 (0.033)		30076.52
model 13		-0.096 (0.052)$^{(\cdot)}$	0.002 (0.005)			-0.04 (0.049)	29636.98
model 14		-0.115 (0.051)$^{(*)}$		0.002 (0.021)	-0.009 (0.04)		30065.45
model 15		-0.102 (0.059)$^{(\cdot)}$		0.008 (0.018)		-0.039 (0.045)	29692.26
model 16		-0.105 (0.057)$^{(\cdot)}$			-0.02 (0.033)	-0.043 (0.044)	29646.01
model 17			-0.003 (0.005)	-0.011 (0.021)	0.001 (0.041)		26685.22

Table 4.7: *The parameter estimates and standard errors of the macro-economic factors in thirty-five different models, along with the AIC_{cd}-values of these models. Each line represents one model. These thirty-five mixture cure models include the seven stated time-independent covariates, but their parameter estimates are omitted here for brevity sake. Significance of the estimates is denoted by $(*)$ (significant at the 5% level) and $(\cdot)$ (significant at the 10% level).*

	interest rate	BEL 20 index	cons confidence	GDP	inflation rate	unemployment	AIC_{cd}
model 18			-0.003 (0.005)		0.005 (0.031)	-0.051 (0.05)	29378.38
model 19				-0.014 (0.023)	-0.003 (0.04)	-0.056 (0.039)	29423.4
model 20			-0.002 (0.005)	-0.009 (0.016)		-0.049 (0.049)	29398.83
model 21	0.068 (0.039)$^{(\cdot)}$	-0.135 (0.063)$^{(*)}$	0.001 (0.005)	0.019 (0.021)			29849.33
model 22	0.073 (0.038)$^{(\cdot)}$	-0.134 (0.06)$^{(*)}$	0 (0.005)		-0.044 (0.038)		29669.83
model 23	0.057 (0.043)	-0.107 (0.054)$^{(*)}$	0.002 (0.005)			0.003 (0.062)	29706.73
model 24	0.071 (0.036)$^{(\cdot)}$	-0.14 (0.057)$^{(*)}$		0.009 (0.022)	-0.033 (0.041)		27743.78
model 25	0.074 (0.047)	-0.138 (0.063)$^{(*)}$		0.02 (0.02)		0.014 (0.057)	29844.81
model 26	0.077 (0.045)$^{(\cdot)}$	-0.137 (0.06)$^{(*)}$			-0.044 (0.036)	0.01 (0.059)	29654.95
model 27	0.052 (0.036)		-0.004 (0.005)	-0.008 (0.022)	-0.015 (0.042)		29449.61
model 28	0.04 (0.043)		-0.003 (0.005)	-0.004 (0.017)		-0.022 (0.058)	29436.45
model 29	0.045 (0.043)		-0.004 (0.005)		-0.007 (0.033)	-0.021 (0.058)	29439.8
model 30	0.037 (0.042)			-0.014 (0.022)	-0.008 (0.04)	-0.033 (0.054)	29478.89
model 31		-0.112 (0.06)$^{(\cdot)}$	-0.001 (0.005)	0.003 (0.022)	-0.011 (0.041)		30089.48
model 32		-0.105 (0.062)$^{(\cdot)}$	0.001 (0.005)	0.006 (0.019)		-0.04 (0.048)	29683.24
model 33		-0.106 (0.059)$^{(\cdot)}$	0 (0.005)		-0.019 (0.035)	-0.044 (0.049)	29644.12
model 34		-0.107 (0.057)$^{(\cdot)}$		0.002 (0.018)	-0.018 (0.04)	-0.043 (0.056)	29659.54
model 35			-0.002 (0.005)	-0.011 (0.022)	-0.008 (0.041)	-0.051 (0.048)	29421.5

Table 4.7 (continued)

ble 4.7, along with a corrected version of the Akaike information criterion (named complete-data AIC or the AIC_{cd}). This AIC_{cd} is based on the converged complete-data log likelihood $Q\big(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}\big)$ instead of the standard log likelihood and can be computed through

$$AIC_{cd} = -2Q\big(\widehat{\boldsymbol{\Theta}} \mid \widehat{\boldsymbol{\Theta}}\big) + 2d + 2\,\text{trace}\{\boldsymbol{DM}(I_d - \boldsymbol{DM})^{-1}\},$$

where d is the length of the parameter vector, I_d is a $d \times d$ identity matrix and DM is the matrix rate of convergence of the EM algorithm, which is automatically computed when using the SEM-algorithm (Meng and Rubin, 1991). The AIC_{cd} in the mixture cure context is discussed in detail in Dirick et al. (2015).

From Table 4.7, it can be seen that the level of the BEL 20 index in general tends to have a significant impact on default, followed by the interest rate in many of the models listed in Table 4.7. The other macro-economic factors; however, do not have a significant effect, which can also be seen through the fact that they tend to switch signs from one model to another. These results are in line with the results shown by Bellotti and Crook (2009), who are also exploring the effect of macro-economic factors on default of individual applicants, using a Cox proportional hazards model. In accordance with this paper, significant effects are observed for the MV interest rate, and both consumer confidence and unemployment rate do not have a significant effect on default. However, the FTSE all-share index, which can be seen as the British equivalent of the BEL 20 index, did not have a significant effect in this study. Bellotti and Crook (2009) found additional significant effects for the UK production index, earnings ratio and the house price index. Figlewski et al. (2012) found that significance and signs of MVs can depend heavily on which other variables are included. In their study, however, corporate default was investigated as opposed to personal loan default.

The full parameter information for the best three models with regard to the AIC_{cd}-values, model 17, 24 and 18, along with the model that only contains the seven time-independent covariates, are given in Table 4.8. As a general result, the AIC_{cd} clearly improves by including TVCs in the

AIC_{cd}		(int)	z_1	z_2	z_3	z_4	z_5	z_6	z_7	IR	BEL 20	cons conf	GDP	infl	unempl
no TVC	***b***	2.492	0.016	0.005	-0.076	-0.023	-0.039	-1.083	-1.319						
		0.015	0.002	0.001	0.033	0.001	0.002	0.038	0.026						
		***	***	***	**	***	***	***	***						
30126.98	***β***		-0.000	-0.011	-0.074	-0.024	-0.023	-0.615	-0.488						
			(0.005)	(0.003)	(0.077)	(0.004)	(0.006)	(0.104)	(0.092)						
			ns	***	ns	***	***	***	***						
model 17	***b***	3.55	0.046	0.058	0.192	0.009	-0.07	-1.406	-2.153						
		(0.085)	(0.002)	(0.002)	(0.037)	(0.002)	(0.002)	(0.04)	(0.033)						
		***	***	***	***	***	***	***	***						
26685.22	***β***		-0.011	-0.026	-0.188	-0.036	-0.016	-0.519	-0.119			-0.003	-0.011	0.001	
			(0.005)	(0.003)	(0.076)	(0.004)	(0.006)	(0.101)	(0.077)			(0.005)	(0.021)	(0.041)	
			*	***	*	***	**	***	ns			ns	ns	ns	
model 24	***b***	2.861	0.035	0.034	0.067	-0.003	-0.059	-1.23	-1.829						
		(0.079)	(0.002)	(0.001)	(0.036)	(0.002)	(0.002)	(0.039)	(0.032)						
		***	***	***	.	*	***	***	***						
27743.78	***β***		-0.008	-0.021	-0.133	-0.031	-0.016	-0.543	-0.188	0.071	-0.14		0.009	-0.033	
			(0.005)	(0.003)	(0.077)	(0.004)	(0.005)	(0.1)	(0.075)	(0.036)	(0.057)		(0.022)	(0.041)	
			.	***	.	***	**	***	*	.	*		ns	ns	
model 18	***b***	3.309	0.028	0.023	0	-0.011	-0.048	-1.22	-1.588						
		(0.072)	(0.002)	(0.001)	(0.034)	(0.002)	(0.002)	(0.037)	(0.033)						
		***	***	***	ns	***	***	***	***						
29378.38	***β***		-0.002	-0.015	-0.108	-0.029	-0.025	-0.659	-0.489			-0.003		0.005	-0.051
			(0.005)	(0.002)	(0.074)	(0.004)	(0.006)	(0.102)	(0.085)			(0.005)		(0.031)	(0.05)
			ns	***	ns	***	***	***	***			ns		ns	ns

Table 4.8: *Full parameter information of the three best models (models 17, 24 and 18 in Table 4.7), out of the thirty-five examined models, according to their AIC_{cd}-values, along with the parameter estimates of the model without TVCs. Significance is denoted by (·) (significant at the* 10% *level), (*) (significant at the* 5% *level), (**) (significant at the* 1% *level) and (***) (significant at the* 0.1% *level).*

models. On the other hand, a lower AIC_{cd} does not guarantee a model with significant TVCs, as can be seen in the AIC_{cd} "best" model 17. For an explanation on how particular parameter estimates affect default, we look at the model with significant effects for the interest rate and BEL 20 index. Residential stability (z_4), length of employment (z_5), a higher bureau score (z_6) and the presence of a mortgage (z_7) lead to a lower susceptibility (through the negative values of estimates $\hat{\boldsymbol{b}}$). The according negative $\hat{\boldsymbol{\beta}}$-estimates for these four variables lead to a longer time until default. Both $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{\beta}}$ indicate that debtors with more job and residential stability, as well as a higher bureau score and with a real estate mortgage tend to be less prone to default. The effect of annual income (z_1), age (z_2) and presence of a monthly child allowance (z_3) is less clear: with positive $\hat{\boldsymbol{b}}$-estimates, susceptibility to default is increased, but the negative $\hat{\boldsymbol{\beta}}$-estimates however indicate delayed default. When looking at the TVCs, logically, a higher interest rate leads to an increase in default hazard ($\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}$ has positive sign), and a better state of the BEL 20 index leads to a decrease in default (negative sign). With insignificant effects of the gross domestic product and inflation rate on default, there is no conclusive effect of these TVCs on default.

4.7.2 Extension: the multiple event mixture cure model

In reality, default is not the only possible event when considering credit risk. Another event type is early repayment, which occurs when a customer repays the loan before the predefined end term. The mixture cure model could be used to repeat the exact same analysis for modeling early repayment instead of default, but it is also possible to include early repayment as an extra term in the mixture cure model (for more information on mixture cure models with multiple events, see Watkins et al., 2014; Dirick et al., 2015). For this type of models, event-specific censoring indicators $\delta_{i,d}$, $\delta_{i,e}$ and $\delta_{i,m}$ are introduced (denoting default, early repayment and maturity indicators respectively), along with a general censoring indicator $\delta_i = \delta_{i,d} + \delta_{i,e} + \delta_{i,m}$ for each observation i. Analogously to the suscepti-

AIC_{cd}	**d**/**e**	(int)	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$	$\hat{b}_6$	$\hat{b}_7$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	IR	BEL20	C Conf	GDP	infl rate	unemp
model 1	**d**	-0.036	0.006	-0.013	-0.102	-0.037	-0.047	-1.245	-1.208	-0.001	-0.001	-0.051	-0.007	-0.006	-0.2	-0.105			0.001		0.027	-0.015
		0.165	0.004	0.003	0.078	0.004	0.005	0.099	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086			0.005		0.033	0.049
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns			ns		ns	ns
98787.8	*e*	0.691	-0.007	-0.017	0.009	-0.017	-0.014	-0.741	-0.218	0.002	-0.001	0.035	-0.002	-0.001	-0.175	-0.134			0.002		0.017	-0.019
		0.061	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.036			0.002		0.016	0.023
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***			ns		ns	ns
model 2	**d**	-0.037	0.006	-0.013	-0.102	-0.037	-0.047	-1.245	-1.208	-0.001	-0.001	-0.05	-0.007	-0.006	-0.199	-0.107		0.014	-0.002			-0.024
		0.164	0.004	0.003	0.078	0.004	0.005	0.099	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086		0.056	0.005			0.052
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns		ns	ns			ns
98767.4	*e*	0.688	-0.007	-0.017	0.009	-0.017	-0.014	-0.74	-0.218	0.002	-0.001	0.036	-0.002	-0.001	-0.177	-0.134		-0.043	0.003			-0.017
		0.061	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.035		0.026	0.002			0.023
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***		ns	ns			ns
model 3	**d**	-0.038	0.006	-0.013	-0.102	-0.037	-0.047	-1.245	-1.208	-0.001	-0.001	-0.051	-0.007	-0.006	-0.201	-0.106		0.021		-0.01		-0.025
		0.156	0.004	0.003	0.078	0.004	0.005	0.099	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086		0.058		0.017		0.047
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns		ns		ns		ns
98791.1	*e*	0.691	-0.007	-0.017	0.009	-0.017	-0.014	-0.741	-0.217	0.002	-0.001	0.036	-0.002	-0.001	-0.177	-0.135		-0.054		0.014		-0.017
		0.024	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.036		0.028		0.008		0.022
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***		.		.		ns
model 4	**d**	-0.034	0.006	-0.013	-0.102	-0.037	-0.047	-1.246	-1.208	-0.001	-0.001	-0.052	-0.007	-0.006	-0.199	-0.103		0.033	-0.001		0.037	
		0.165	0.004	0.003	0.078	0.004	0.005	0.099	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086		0.059	0.005		0.035	
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns		ns	ns		ns	
98785.5	*e*	0.69	-0.007	-0.017	0.009	-0.017	-0.014	-0.741	-0.218	0.002	-0.001	0.036	-0.002	-0.001	-0.177	-0.135		-0.039	0.003		0.011	
		0.062	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.036		0.028	0.002		0.017	
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***		ns	ns		ns	
model 5	**d**	-0.036	0.006	-0.013	-0.103	-0.037	-0.047	-1.245	-1.208	-0.001	-0.001	-0.046	-0.007	-0.006	-0.199	-0.109	0.046		-0.001	0.002		
		0.167	0.004	0.003	0.078	0.004	0.005	0.099	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086	0.036		0.005	0.018		
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns	ns		ns	ns		
98798.2	*e*	0.692	-0.007	-0.017	0.009	-0.017	-0.014	-0.741	-0.218	0.002	-0.001	0.035	-0.002	-0.001	-0.174	-0.134	0.007		-0.001	0.003		
		0.065	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.036	0.017		0.002	0.008		
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***	ns		ns	ns		

Table 4.9: *The parameter estimates of ten multiple event mixture cure models containing TVCs.* **d** *are parameter estimates related to the default event,* **e** *denotes early repayment parameter estimates.*

AIC_{cd}	d/e	(int)	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$	$\hat{b}_6$	$\hat{b}_7$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	IR	BEL20	C Conf	GDP	infl rate	unemp
model 6	***d***	-0.036	0.006	-0.013	-0.103	-0.037	-0.047	-1.245	-1.208	-0.001	-0.001	-0.046	-0.007	-0.006	-0.199	-0.109	0.046	0.005	-0.001	0.001		
		0.168	0.004	0.003	0.078	0.004	0.005	0.099	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086	0.037	0.063	0.005	0.02		
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns	ns	ns	ns	ns		
98791.3	***e***	0.691	-0.007	-0.017	0.009	-0.017	-0.014	-0.741	-0.218	0.002	-0.001	0.037	-0.002	-0.001	-0.177	-0.134	0.016	-0.068	0.001	0.014		
		0.066	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.036	0.017	0.03	0.002	0.009		
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***	ns	*	ns	ns		
model 7	***d***	-0.033	0.006	-0.014	-0.103	-0.037	-0.047	-1.246	-1.208	-0.001	-0.001	-0.049	-0.007	-0.006	-0.198	-0.107	0.039	0.011		0.007	0.029	
		0.163	0.004	0.003	0.078	0.004	0.005	0.1	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086	0.038	0.061		0.022	0.042	
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns	ns	ns		ns	ns	
98836.3	***e***	0.695	-0.007	-0.017	0.008	-0.017	-0.014	-0.741	-0.217	0.002	-0.001	0.037	-0.002	-0.001	-0.176	-0.136	0.01	-0.056		0.023	0.022	
		0.095	0.002	0.001	0.034	0.002	0.002	0.041	0.034	0.002	0.001	0.036	0.002	0.002	0.042	0.035	0.019	0.025		0.009	0.023	
		***	**	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***	ns	*		*	ns	
model 8	***d***	-0.037	0.006	-0.013	-0.103	-0.037	-0.047	-1.245	-1.208	-0.001	-0.001	-0.047	-0.007	-0.006	-0.2	-0.109	0.046	-0.001		-0.001		
		0.175	0.004	0.003	0.078	0.004	0.005	0.099	0.083	0.006	0.003	0.079	0.004	0.006	0.103	0.086	0.036	0.057		0.019		
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns	ns	ns		ns		
98796.6	***e***	0.692	-0.007	-0.017	0.009	-0.017	-0.014	-0.741	-0.217	0.002	-0.001	0.037	-0.002	-0.001	-0.177	-0.134	0.016	-0.064		0.016		
		0.076	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.035	0.017	0.027		0.009		
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***	ns	*		.		
model 9	***d***	-0.038	0.006	-0.013	-0.102	-0.037	-0.047	-1.245	-1.208	-0.001	-0.001	-0.047	-0.007	-0.006	-0.2	-0.108	0.052			-0.002		0.013
		0.175	0.004	0.003	0.078	0.004	0.005	0.099	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086	0.043			0.014		0.052
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns	ns			ns		ns
98781.6	***e***	0.69	-0.007	-0.017	0.009	-0.017	-0.014	-0.741	-0.218	0.002	-0.001	0.034	-0.002	0	-0.175	-0.134	-0.003			0.003		-0.026
		0.077	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.035	0.019			0.007		0.028
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***	ns			ns		ns
model 10	***d***	-0.037	0.006	-0.013	-0.103	-0.037	-0.047	-1.245	-1.208	-0.001	-0.001	-0.047	-0.007	-0.006	-0.2	-0.108	0.054	-0.007				0.016
		0.175	0.005	0.003	0.078	0.004	0.005	0.099	0.083	0.005	0.003	0.079	0.004	0.006	0.103	0.086	0.042	0.044				0.061
		ns	ns	***	ns	***	***	***	***	ns	ns	ns	.	ns	.	ns	ns	ns				ns
98795.8	***e***	0.692	-0.007	-0.017	0.009	-0.017	-0.014	-0.741	-0.217	0.002	-0.001	0.036	-0.002	-0.001	-0.177	-0.135	-0.002	-0.024				-0.013
		0.077	0.002	0.001	0.034	0.002	0.002	0.04	0.035	0.002	0.001	0.035	0.002	0.002	0.042	0.035	0.02	0.021				0.028
		***	***	***	ns	***	***	***	***	ns	ns	ns	ns	ns	***	***	ns	ns				ns

Table 4.9 (continued)

bility indicator Y_i, three indicators $Y_{i,d}, Y_{i,e}$ and $Y_{i,m}$ are introduced. The unconditional survival function of the multiple event mixture cure model for given $\boldsymbol{z} and \boldsymbol{x}(t)$ is then given by

$$\begin{aligned} S(t \mid \boldsymbol{z}, \boldsymbol{x}(t)) &= \pi_e(\boldsymbol{z}) S_e(t \mid Y_e = 1; \boldsymbol{z}, \boldsymbol{x}(t)) \\ &\quad + \pi_d(\boldsymbol{z}) S_d(t \mid Y_d = 1; \boldsymbol{z}, \boldsymbol{x}(t)) \\ &\quad + \big(1 - \pi_e(\boldsymbol{z}) - \pi_d(\boldsymbol{z})\big), \end{aligned}$$

with $S_e(t \mid Y_e = 1, \boldsymbol{z}, \boldsymbol{x}(t))$ and $S_d(t \mid Y_d = 1; \boldsymbol{z}, \boldsymbol{x}(t))$ the conditional survival functions for early repayment and default respectively. These functions are modeled using two Cox PH models, as in (4.5).

Two major changes with regard to the single event mixture cure model are the computation of $\pi_d(\boldsymbol{z})$ and $\pi_e(\boldsymbol{z})$, and the conditional expectations of the Y-indicators, resulting in the weights w. With more than two groups, the binomial logit is replaced by the multinomial logit,

$$\pi_d(\boldsymbol{z}) = P(Y_d = 1; \boldsymbol{z}) = \frac{\exp(b_d{}' \boldsymbol{z})}{1 + \exp(b_d{}' \boldsymbol{z}) + \exp(b_e{}' \boldsymbol{z})}.$$

$\pi_e(\boldsymbol{x})$ is found analogously. As an extension to (4.10), the event-specific weights for early repayment and default can be computed, with in this case $\boldsymbol{\Theta} = (\boldsymbol{b}, \boldsymbol{\beta}_d, \boldsymbol{\beta}_{t,d}, \boldsymbol{\beta}_e, \boldsymbol{\beta}_{t,e})$, and $\mathbf{O} = (\lambda_{d,i,j}, \lambda_{e,i,j}, \delta_i, \delta_{i,d}, \delta_{i,e}, \delta_{i,m}, t_{i,d}, t_{i,e})$. The interval-specific censoring indicators λ as well as the event time t depend on the event type, default or early repayment. The event-specific weight for default is then given by (covariates are omitted in the arguments of S_d and S_e for brevity sake)

$$w_{i,d}^{(r)} = E(Y_{i,d} \mid \mathbf{O}, \boldsymbol{\Theta}^{(r)}) \tag{4.12}$$

$$= \begin{cases} \dfrac{\pi_d(\boldsymbol{z}_i) S_d(t_i \mid Y_{i,d} = 1)}{\pi_e(\boldsymbol{z}_i) S_e(t_i \mid Y_{i,e} = 1) + \pi_d(\boldsymbol{z}_i) S_d(t_i \mid Y_{i,d} = 1) + \big(1 - \pi_e(\boldsymbol{z}_i) - \pi_d(\boldsymbol{z}_i)\big)} & \text{for } \delta_i = 0 \\ 1 & \text{for } \delta_{i,d} = 1. \\ 0 & \text{for } \delta_{i,d} = 0\,; \delta_i = 1 \end{cases}$$

Note that, when $\delta_i = 0$, $t_{i,d} = t_{i,e} = t_i$, $w_{i,e}^{(r)}$ can be computed in a similar fashion. Again, the EM-algorithm is used for computation of the expected complete-data log likelihood.

The multiple event mixture cure model was applied to the data, adding the additional information of early repayments, which was ignored when

applying the single event mixture cure model. The result of ten arbitrarily selected models was listed in Table 4.9. In each model, three to four TVCs were present for each event-type, each time including the same TVC for the default and early repayment events. If we first consider the parameter estimates of the default events, we notice that, where the effect and significance of the $\hat{b}$-parameter estimates on default lie relatively close to the results in Table 4.8, nearly all $\hat{\beta}$-parameter estimates became insignificant. The significant effects that remain, are the number of years at current address and the bureau score. Additionally, no significant $\hat{\beta}_t$-parameter effect remains for default. Looking at the early repayment parameter estimates, a notable remark is that the signs of the parameters here tend to be the same as the signs for the default parameter. While an early repayment does not immediately incur costs for a bank, this event type does lead to a decline in expected revenue, as the interest payments for the months following the time of early repayment are lost. In fact, one can look at both default and early repayment as events that are results of a common trigger, which is customer instability. Therefore early repayment must be seen as a negative event that banks prefer to avoid, and this is also reflected in the parameter estimates. For early repayment, two TVCs tend to have a significant effect on the hazard of early repayment: the BEL20 index and the gross domestic product.

4.7.3 Extension: The mixture cure model with piecewise linear relationship for the TVCs

With abundant information on the TVCs (with one TVC-value per subject per month that the subject is observed), estimating just one $\hat{\beta}_t$-parameter estimate for each TVC might be overly simplistic. On the other hand, the effect of a certain TVC-value on default might depend on the specific range this TVC-value is in. For example, the effect of the TVC associated with the GDP-value might be different when the GDP is declining with respect to the previous year, compared with when GDP is increasing. A way of overcoming this is by using piecewise linear functions instead of just one

linear effect (or one $\hat{\beta}_t$) per TVC. When constructing a model, the choice between a piecewise constant versus a piecewise linear relationship should be carefully considered.

Applied to our data, six models were created, with each time the seven "baseline" time-independent covariates and one of the TVCs, split up into four piecewise linear functions. The TVC-part of the linear predictor in (4.5), $\boldsymbol{\beta}'\boldsymbol{z} + \boldsymbol{\beta}_t'\boldsymbol{x}(t)$, is then replaced by (for $j = 1, \ldots, 6$, as there are six models with each time one TVC $\bar{x}_j(t)$):

$$\beta_{t1}\bar{x}_j(t)+\beta_{t2}\big(\bar{x}_j(t) - Q_1\big)_+ + \beta_{t3}\big(\bar{x}_j(t) - Q_2\big)_+ + \beta_{t4}\big(\bar{x}_j(t) - Q_3\big)_+, \quad (4.13)$$

where Q_1, Q_2 and Q_3 refer to the first quantile, the second quantile (or median value) and the third quantile of all the TVC-values of the relevant macro-economic factor in the data set. The notation $\big(\bar{x}_j(t) - Q_*\big)_+$ denotes that for each observation and time-period, either $\bar{x}_j(t) - Q_*$ or 0 is retained as a regressor, the latter one when $\bar{x}_j(t) - Q_* < 0$. The result of this construction is that the effect of a TVC changes depending on whether the $x_j(t)$ lies in the interval $[0, Q_1]$, $[Q_1, Q_2]$,$[Q_2, Q_3]$ or $[Q_3, Q_4]$ (respectively β_{t1}, $\beta_{t1} + \beta_{t2}$, $\beta_{t1} + \beta_{t2} + \beta_{t3}$ and $\beta_{t1} + \beta_{t2} + \beta_{t3} + \beta_{t4}$).

In Table 4.10, six models can be found where each one contains one of the six TVCs described in Table 4.6. Consider the first model, with TVC interest rate. While the effect between interest rate and default hazard rate is negative in the interval $[0, Q_1]$, the effect is positive (as would be expected) in all other intervals (the effects are -0.064, 0.404, 0.343 and 0.034 respectively). In any case, we see here that the effect of the middle ranges of the interest rate seems to be more distinct (0.404, 0.343) compared to the "border intervals". However, it should be noted that from the results, no conclusions can be drawn, as none of the estimates are significant. This is in general the case in Table 4.10.

4.8 Discussion

We showed that, as an addition to survival analysis, time-dependent covariates can be included in mixture cure models. The AIC_{cd} showed that

TVC in model		(int)	z_1	z_2	z_3	z_4	z_5	z_6	z_7	$\hat{\beta}_{t1}$	$\hat{\beta}_{t2}$	$\hat{\beta}_{t3}$	$\hat{\beta}_{t4}$
Interest rate	$\boldsymbol{b}$	4.019	0.05	0.041	0.062	0.005	-0.066	-1.506	-1.95				
		0.084	0.002	0.002	0.035	0.002	0.002	0.039	0.034				
		***	***	***	.	**	***	***	***				
AIC_{cd}= 27484.3	$\boldsymbol{\beta}$		-0.007	-0.02	-0.131	-0.033	-0.02	-0.574	-0.351	-0.064	0.468	-0.061	-0.309
			0.004	0.003	0.079	0.003	0.006	0.101	0.083	0.057	0.344	0.509	0.39
			.	***	.	***	***	***	***	ns	ns	ns	ns
BEL 20 index	$\boldsymbol{b}$	2.766	0.029	0.011	0	-0.012	-0.043	-1.145	-1.483				
		0.072	0.002	0.001	0.034	0.002	0.002	0.037	0.033				
		***	***	***	ns	***	***	***	***				
AIC_{cd}=29781.1	$\boldsymbol{\beta}$		-0.004	-0.013	-0.108	-0.028	-0.023	-0.619	-0.429	-0.021	-0.132	-0.799	2.202
			0.003	0.004	0.08	0.002	0.006	0.102	0.078	0.086	0.193	0.551	0.932
			ns	**	ns	***	***	***	***	ns	ns	ns	*
Cons confidence	$\boldsymbol{b}$	3.289	0.029	0.024	0.014	-0.01	-0.049	-1.221	-1.607				
		0.073	0.002	0.001	0.034	0.002	0.002	0.038	0.033				
		***	***	***	ns	***	***	***	***				
AIC_{cd}=29303.8	$\boldsymbol{\beta}$		-0.003	-0.016	-0.111	-0.03	-0.025	-0.654	-0.472	-0.007	0.001	0	0.003
			0.005	0.002	0.077	0.004	0.006	0.102	0.086	0.014	0.025	0.03	0.033
			ns	***	ns	***	***	***	***	ns	ns	ns	ns
GDP	$\boldsymbol{b}$	2.964	0.025	0.023	-0.008	-0.012	-0.047	-1.142	-1.617				
		0.072	0.002	0.001	0.034	0.002	0.002	0.038	0.032				
		***	***	***	ns	***	***	***	***				
AIC_{cd}=29226.1	$\boldsymbol{\beta}$		-0.002	-0.016	-0.1	-0.029	-0.023	-0.649	-0.404	-0.051	0.158	-0.181	0.053
			0.005	0.002	0.075	0.004	0.006	0.102	0.084	0.031	0.148	0.242	0.14
			ns	***	ns	***	***	***	***	.	ns	ns	ns
Inflation rate	$\boldsymbol{b}$	2.824	0.024	0.019	0	-0.014	-0.045	-1.139	-1.556				
		0.072	0.002	0.001	0.034	0.002	0.002	0.038	0.032				
		***	***	***	ns	***	***	***	***				
AIC_{cd}=29441.7	$\boldsymbol{\beta}$		-0.003	-0.015	-0.104	-0.028	-0.023	-0.626	-0.405	0.077	-0.269	0.37	-0.131
			0.006	0.002	0.071	0.005	0.006	0.102	0.09	0.066	0.184	0.269	0.183
			ns	***	ns	***	***	***	***	ns	ns	ns	ns
Unemployment	$\boldsymbol{b}$	3.342	0.03	0.021	0.003	-0.01	-0.047	-1.246	-1.55				
		0.073	0.002	0.001	0.034	0.002	0.002	0.037	0.033				
		***	***	***	ns	***	***	***	***				
AIC_{cd}=29500.7	$\boldsymbol{\beta}$		-0.003	-0.015	-0.109	-0.03	-0.025	-0.649	-0.506	0.007	-0.296	0.656	-0.623
			0.005	0.002	0.073	0.004	0.006	0.102	0.088	0.169	0.283	0.431	0.444
			ns	***	ns	***	***	***	***	ns	ns	ns	ns

Table 4.10: *Six mixture cure models, containing one TVC each time split up into four piecewise linear pieces bounded by the quantiles of the TVC of interest, as denoted in (4.13).*

including TVCs can indeed lead to a better model fit. Using piecewise linear functions, more complex relationships between the TVCs and the event of interest can be modeled. Extending the mixture cure model to multiple events, both default and early repayment events can be joined in one model.

A general result for the data sets we have used in this study is that

only a limited number of macro-economic factors tended to have an effect on default (in the single event mixture cure model) and early repayment (in the multiple event mixture cure model). Where the BEL20 index had an influence on both event types, the interest rate had an influence on the former and GDP on the latter event type only. It is indeed plausible that some macro-economic factors do not affect default or early repayment. Let us take the unemployment rate as an example: because of a selection bias (as banks only granted loans to supposedly creditworthy customers), the debtors in the data set might not be affected by higher unemployment, if a rise in unemployment was not present among the subjects in the data set. On the other hand, some actual effects of TVC on default might be lost as a result of the averaging of TVCs over a monthly period. Interesting results might be obtained by applying the models when looking at weekly or even daily TVC levels; however, this would largely increase the data size, and requires daily information of default and early repayment events.

For future research, going deeper into the piecewise linear modeling of the TVCs might be advisable. An interesting research focus here is finding appropriate "knots", which were chosen to be the quantile values in this chapter. The choice of different knots may lead to more insightful results.

Chapter 5

General conclusions and research perspectives

In the context of selecting an appropriate survival analysis technique in the credit risk field, this thesis investigated and extended different aspects of mixture cure models. This final chapter concludes this work by giving some suggestions for future research.

The first chapter of this thesis focused on finding an appropriate version of the AIC, using the expected complete-data log likelihood as a result of the EM algorithm. Emphasizing other aspects of the modeling procedure would lead to the development of other selection methods. A Bayesian information criterion for these models is expected to have consistency properties, however, under the strong (and unrealistic) assumption that the true credit risk model is exactly described by one of the used models. A focused information criterion (Claeskens and Hjort, 2003) would rather assume local misspecification and selects a model that is best in terms of mean squared error or mean squared prediction error for a certain focus quantity (such as the probability of the time to default to fall in a certain period).

In chapter 2, unobserved heterogeneity is modeled through explicit inclusion of separate survival contributions, depending on the "group" that one belongs to. Another possible approach could be trough the use of

frailty models. In these models, the unobserved heterogeneity is modeled through a random variable that describes unobserved risk factors. The advantage of frailty models is that they allow for more flexibility in comparison with the model proposed in chapter 2. From a practitioner's point of view, however, an important drawback is that these models are more complex in terms of interpretation. For more information on frailty models, we refer to Hougaard (1995); Gutierrez (2002); Duchateau and Janssen (2008).

The need for an appropriate evaluation measure to assess the performance of survival methods becomes apparent in chapter 3. As a result from the extensive use of classification techniques in the banking sectors, ROC-curves and their AUCs are still widespread means of performance estimation. Comparing the expected future value using the models and the actual future value, as done in chapter 3, is a step in the right direction, but more work needs to be done as a substantial part of the model performance on the data is still not evaluated due to censoring.

As already noted in chapter 3, more research on piecewise linear modeling of the TVCs would be desirable in order to get a better idea on how macro-economic variables affect loan default. Additionally, it should be noted that the implementation of the TVCs can be extended even further, as currently TVCs can only be included in the latency part of the mixture cure model. Including TVCs in the incidence model, a binomial logistic regression model, is not possible as such. Additional research could give interesting insights. This extension is possibly not feasible in mixture cure models, and the shift to a related but different type of cure model, the promotion time cure model (Yakovlev et al., 1993), is needed to achieve this. As a third research prospective related to this chapter, we mention the extension of mixture cure models such that not only external but also internal TVCs can be included.

Having worked on the topic for four years, it needs to be said that banks are still hesitant to use survival analysis models, notwithstanding the extensive research and the promising results. The reason for this hesitation is very simple: fear of model complexity in comparison with classification

techniques. I sincerely hope that through my presentations at banks and practitioners conferences, and this thesis, I have contributed in paving the way for more use of survival analysis techniques in the banking sector.

List of Figures

List of Tables

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Andersen, P. K. (1992). Repeated assessment of risk factors in survival analysis. *Statistical Methods in Medical Research*, 1(3):297–315.

Andreeva, G. (2006). European generic scoring models using survival analysis. *The Journal of the Operational Research Society*, 57(10):1180–1187.

Austin, P. C. (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state of the art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635.

Banasik, J., Crook, J., and Thomas, L. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, 50(12):1185–1190.

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society*, 60(12):1699–1707.

Berrington, A. and Diamond, I. (2000). Marriage or cohabitation: a competing risks analysis of first-partnership formation among the 1958 british birth cohort. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(2):127–151.

Bremhorst, V. and Lambert, P. (2014). Flexible estimation in cure survival models using Bayesian P-splines. *Computational Statistics & Data Analysis*, pages 1 –15. in press.

Burda, M., Harding, M., and Hausman, J. (2015). A bayesian semiparametric competing risk model with unobserved heterogeneity. *Journal of Applied Econometrics*, 30(3):353–376.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012a). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108:1255–1260.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012b). *smcure: Fit Semiparametric Mixture Cure Models*. R package version 2.0.

Cai, L. and Lee, T. (2009). Covariance structure model fit testing under missing data: An application of the supplemented EM algorithm. *Multivariate Behavioral Research*, 44(2):281–304.

Cao, R., Vilar, J. M., and Devia, A. (2009). Modelling consumer credit risk via survival analysis. *SORT*, 33(1):3–30.

Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference*, 67(1):45–65.

Ciochetti, D., Deng, Y., Gao, B., and Yao, R. (2002). The termination of commercial mortgage contracts through prepayment and default: A proportional hazards approach with competing risks. *Real Estate Economics*, 30(4):595–633.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, 64:1062–1069.

Claeskens, G. and Hjort, N. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900–916. With discussion and a rejoinder by the authors.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Collett, D. (2003). *Modelling Survival Data in Medical Research, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Cortese, G. and Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal*, 52(1):138–158.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

de Boor, C. (2001). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer New York.

Dejaeger, K., Verbeke, W., Martens, D., and Baesens, B. (2012). Data mining techniques for software effort estimation: A comparative study. *Software Engineering, IEEE Transactions on*, 38(2):375–397.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Deng, Y., Quigley, J., and Van Order, R. (2000). Mortgage terminations, heterogeneity, and the exercise of mortgage options. *Econometrica*, 68(2):275–307.

Dirick, L., Claeskens, G., and Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, 241:449–457.

Donohue, M., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, 98(3):685–700.

Duchateau, L. and Janssen, P. (2008). *The frailty model*. Statistics for Biology and Health series. Springer Verlag.

Eilers, P. H. C., Rijnmond, D. M., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, 30:74–99.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046.

Figlewski, S., Frydman, H., and Liang, W. (2012). Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics & Finance*, 21(1):87–105.

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157.

Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*.

Fox, J. and Carvalho, M. S. (2012). The RcmdrPlugin.survival package: Extending the R commander interface to survival analysis. *Journal of Statistical Software*, 49(7):1–32.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Gutierrez, R. G. (2002). Parametric frailty and shared frailty survival models. *Stata Journal*, 2(1):22–44.

Heagerty, P. and Saha, P. (2000). SurvivalROC: time-dependent roc curve estimation from censored survival data. *Biometrics*, 56(2):337–344.

Heckman, J. J. and Honoré, B. E. (1989). The identifiability of the competing risks model. *Biometrika*, 76(2):325–330.

Hjort, N. and Claeskens, G. (2006). Focussed information criteria and model averaging for Cox's hazard regression model. *Journal of the American Statistical Association*, 101:1449–1464.

Hosmer, D. W., May, S., and Lemeshow, S. (2008). *Applied Survival Analysis: Regression Modelling of Time to Event Data*. Wiley-Interscience, second edition.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, 1(3):255–273.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, 103(484):1648–1658.

Jamshidian, M. and Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):257–270.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New Jersey, 2 edition.

Kellison, S. G. and Irwin, R. D. (1991). *The Theory Of Interest*, volume 2. Irwin Homewood, IL.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer.

Kleinbaum, D. and Klein, M. (2011). *Survival Analysis: A Self-Learning Text, Third Edition*. Statistics for Biology and Health. Springer.

Kuk, A. and Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541.

Liang, H. and Zou, G. (2008). Improved AIC selection strategy for survival analysis. *Computational Statistics & Data Analysis*, 52(5):2538 – 2548.

Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics*, 51(2):524–532.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

Naik, P. A., Shi, P., and Tsai, C.-L. (2007). Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102(477):244–254.

Narain, B. (1992). Survival analysis and the credit granting decision. In Thomas, L. C., Crook, J. N., and Edelman, D. B., editors, *Credit Scoring and Credit Control*, pages 109–121. Clarendon Press, Oxford.

NBB (2015). National bank of belgium: Database with macro-economic factors. `http://stat.nbb.be`.

Ng, A. K., Bernardo, M. P., Weller, E., Backstrand, K. H., Silver, B., Marcus, K. C., Tarbell, N. J., Friedberg, J., Canellos, G. P., and Mauch, P. M. (2002). Long-term survival and competing causes of death in patients with early-stage Hodgkińs disease treated at age 50 or younger. *Journal of Clinical Oncology*, 20(8):2101–2108.

O'Sullivan, F. et al. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518.

Pavlov, A. (2001). Competing risks of mortgage termination: Who refinances, who moves and who defaults. *Journal of Real Estate Economics and Finance*, 23(2):185–211.

Peng, Y. (2003). Fitting semiparametric cure models. *Computational Statistics & Data Analysis*, 41(3-4):481 – 490.

Peng, Y. and Dear, K. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):227–236.

Peng, Y. and Zhang, J. (2008). Identifiability of a mixture cure frailty model. *Statistics & Probability Letters*, 78(16):2604 – 2608.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge university press.

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*, 53(4):1–37.

Segal, M. R., Bacchetti, P., and Jewell, N. P. (1994). Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):345–352.

Stepanova, M. and Thomas, L. (2001). PHAB scores - proportional hazards analysis behavioural scores. *The Journal of the Operational Research Society*, 41(9):1007–1016.

Stepanova, M. and Thomas, L. (2002a). Survival analysis methods for personal loan data. *Operations Research*, 50(2):277–289.

Stepanova, M. and Thomas, L. (2002b). Survival analysis methods for personal loan data. *Operations Research Quarterly*, 50(2):277–289.

Suzukawa, A., Imai, H., and Sato, Y. (2001). Kullback-Leibler information consistent estimation for censored data. *Annals of the Institute of Statistical Mathematics*, 53(2):262–276.

Sy, J. and Taylor, J. (2000a). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Sy, J. P. and Taylor, J. M. G. (2000b). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Therneau, T. M. (2014). *A Package for Survival Analysis in S*. R package version 2.37-7.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Thomas, L., Edelman, D., and Crook, J. (2002). *Credit Scoring and Its Applications*. Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics.

Thomas, L. C. (2009). *Consumer Credit Models: Pricing, Profit and Portfolios: Pricing, Profit and Portfolios*. OUP Oxford.

Tong, E. N. C., Mues, C., and Thomas, L. C. (2012). Mixture cure models in credit scoring: if and when borrowers default. *European Journal of Operational Research*, 218(1):132–139.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

Van Gestel, T. and Baesens, B. (2008). *Credit Risk Management : Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. OUP Oxford.

Watkins, J. G. T., Vasnev, A. L., and Gerlach, R. (2014). Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. *Journal of Applied Econometrics*, 29:627–648.

Xu, R., Vaida, F., and Harrington, D. P. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica*, 19(2):819–842.

Yakovlev, A. Y., Asselain, B., Bardou, V.-J., Fourquet, A., Hoang, T., A., R., and Tsodikov, A. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. In Asselain, B., Boniface, M., Duby, C., Lopez, C., Masson, J., and Tranchefort, J., editors, *Biometrie et Analyse de Donnees Spatio-Temporelles*, volume 12, pages 66–82. Rennes, France.

Zhang, J. and Thomas, L. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, 18(2):204–215.

Doctoral dissertations of the Faculty of Economics and Business

A list of doctoral dissertations from the Faculty of Economics and Business can be found at the following website:
`http://www.kuleuven.be/doctoraatsverdediging/archief.htm`.